UNIVERSIDAD CARLOS III DE MADRID

TRABAJO FIN DE GRADO

Modelado de Tópicos para perfilado de Blogs

Autor:

Jorge Vázquez Marcos

Supervisor:

Jesús Cid Sueiro

Grado en Ingeniería en Sistemas Audiovisulaes

Departamento de Teoría de la Señal y Comunicaciones

Leganés, Octubre 2017

Resumen:

En este trabajo se analiza el funcionamiento del Modelo de Tópicos Latent Dirichelt Allocation como herramienta para caracterizar una colecciones de Blogs. Lo que se plantea es el estudio de la composición de los tópicos latente descubiertos y la representación de documentos mediante los mismos. Para ello, se proponen visualizaciones de tópicos y visualización de grafos que expresen relaciones entre documentos.

Abstract:

This project studies the Latent Dirichelt Allocation Topic Modelling as a tool to characterize a collection of Blogs. The main goal is to analyze the composition of latent topics discovered and how documents are represented through them. For this purpose, topics visualizations and graphs that express relationships between documents, will form the analysis resources.

Capítulo 1

1. Introducción

Este primer capítulo pretende introducir al lector en el proyecto de investigación que se desarrolla en la presente memoria. Se presentan tanto el marco tecnológico como el socio-económico, ofreciendo así una visión general del entorno que subyace en el campo de la investigación. Además, se establecen las motivaciones y objetivos del proyecto, expuestos atendiendo a la situación actual en el campo de la investigación.

Por último, se establecen la consideraciones más relevantes del marco regulatorio aplicable a la investigación.

1.1. Marco tecnológico

Los avances tecnológicos de las últimas décadas han transformado de forma significativa la forma en los que los seres humanos nos relacionamos y organizamos. En consecuencia, la sociedad se ha visto involucrada en un proceso de cambio a todos los niveles. Estos avances tecnológicos se basan, esencialmente, en las Tecnologías de la Información y Comunicación[1].

Las Tecnologías de la Información y Comunicación, como otros grandes avances ya hicieran en el pasado, nos abren la puerta a una nueva etapa de desarrollo social, la Sociedad de la Información.

¿Qué es la Sociedad de la Información?

La Sociedad de la Información es un nuevo estadio de desarrollo en el que la información de cualquier naturaleza, su generación, su distribución y su uso se han convertido en el eje fundamental de todo lo que hacen los seres humanos[1].

La información y su gestión

La Sociedad de la Información no solo está fundamentada en los avances en materia de telecomunicaciones, tecnologías como TCP/IP o las redes móviles y avances en la transmisión de señal como la Fibra Óptica, han modelado las formas de comunicación en el mundo actual. Sin embargo, estos avances solo constituyen la estructura primaria, es decir, las vías de transmisión de información que han permitido que se desarrolle el mundo interconectado globalmente que hoy conocemos. Por encima de esta estructura primaria se sitúan todos los mecanismo de almacenamiento, gestión, entendimiento y categorizado de la información.

Por ejemplo, la World Wide Web es un sistema de distribución de documentos de hipertexto interconectados y accesibles vía Internet[3]. Toda esa gran cantidad de datos almacenados tienen la finalidad de responder a necesidades de información de los usuarios. Para poder satisfacer estas necesidades son fundamentales los mecanismos de búsqueda y consulta. Es aquí donde entran en juego los motores de búsqueda como Google, que son capaces de indexar resultados ordenados ante una consulta, dando así sentido al proceso de almacenamiento y distribución.

Es en este punto donde se juntan las telecomunicaciones con las ciencias de procesamiento de información, la ingeniería de los sistemas de información y las ciencias de la computación. El desarrollo del sector TIC ha conseguido establecer una red digital, global y estructurada de contenidos y servicios, que conocemos comúnmente como Internet.

A su vez, desde la perspectiva de la Sociedad de la Información, la red de Internet es una fuente de información desestructurada cuyo procesado, gestión y análisis puede revertir en positivo en muchos sectores de la sociedad.

Atendiendo a su definición, la información está constituida por un grupo de datos ya supervisados y ordenados, que sirven para construir un mensaje

basado en un cierto fenómeno o ente. Su aprovechamiento racional es la base del conocimiento[4] .

Ahora bien, ¿qué tipos de información encontramos en Internet?

- Texto (Documentos)
- Documentos estructurados (p.ej. XML)
- Imágenes
- Audio
- Video
- Código fuente
- Aplicaciones, servicios web

Imaginemos ahora un Blog de cocina, por ejemplo, en el se publican recetas presentadas en formato post. Atendiendo a la definición anterior de información, este Blog contendría información culinaria, estando ésta disponible para los usuarios que lo requiriesen.

Sin embargo, la red de Internet contiene una gran variedad de Blogs de diferentes temáticas y escritos en diferentes idiomas. Cada Blog de la red de Internet compone una unidad de información, pero desde una perspectiva global son datos desestructurados con un gran potencial de generar nueva información y conocimiento.

Con este ejemplo concreto, podríamos hablar de clasificación automática de textos, aplicado sobre conjuntos blogs de cocina con la intención de categorizarlos por idioma, tipo de comida o ingredientes necesarios. Pero además, en la mayoría de blogs, los usuarios pueden interactuar dejando comentarios o valoraciones. Toda esa información desestructurada puede gestionarse y procesarse también para establecer un sistema de recomendación de recetas para los usuarios, en función de qué recetas le han gustado.

De la misma manera las plataformas de música en streaming o vídeo bajo demanda, emplean la información de uso de sus usuarios para establecer sistemas de recomendación. Las redes sociales como Twitter o blogs de opinión, proporcionan datos desestructurados que pueden tratarse para realizar análisis de opinión o tendencias.

La red de Internet genera de manera continuada grandes cantidades de información desestructurada, su estructuración y gestión pueden aportar vías hacia el conocimiento en multitud de disciplinas. Por este motivo, la información se ha convertido en el eje fundamental de la sociedad actual.

La investigación y desarrollo de técnicas de obtención, recuperación y procesado de información son una de las puntas de lanza del sector TIC en la actualidad.

1.2. Marco socio-económico

El sector de las telecomunicaciones lleva unos años inmerso en una situación de cambios en el mercado y en su entorno específico. El modelo de negocio tradicional de las operadoras de telecomunicaciones aunaba la oferta de infraestructura de redes para conectividad y el establecimiento de servicios sobre las mismas. Estos servicios desplegados sobre la infraestructura de red eran típicamente la telefonía fija, móvil y recientemente acceso Internet.

Desde la liberalización del sector, se estableció un mercado regulado que permitía la convivencia competitiva de entidades empresariales. Esta competencia se basaba en el modelo de negocio ya mencionado de redes y servicios.

Se ha conformado lo que se conoce como El Nuevo Ecosistema Digital, donde conviven las empresas tradicionales del sector con nuevas unidades empresariales que proveen servicios sobre la infraestructura de red.

Servicios Over The Top

Las empresas de servicios OTT(Over the Top), son proveedoras de servicios de todo tipo sobre Internet. Se denominan Over The Top porque son servicios ofrecidos sobre la infraestructura de red IP, es decir, este tipo de empresa no requiere de inversión en infraestructura para implantar sus productos. Este tipo de empresas basan su modelo de negocio en proporcionar servicios y contenidos. Es decir, hacer negocio con los usuarios finales a través de la conectividad de banda ancha proporcionada por los operadores de telecomunicaciones.

Por tanto los agentes OTT son tecnologías y servicios que se corresponden con los siguientes eslabones de la cadena de valor de Internet:

- Propietarios de los derechos sobre los contenidos:
- Servicios en línea
- Tecnologías habilitadoras
- Interfaz de usuario

La aparición de los servicios OTT, supone la llegada de un nuevo agente económico que ha revolucionado actividades fundamentales para el ser humano como el trabajo, educación, entretenimiento, comercio o las relaciones sociales.

En el uso de estos servicios, los usuarios generan grandes cantidades de información que pueden resultar de interés. Bien ya sea para personalizar la experiencia del usuario en el servicio, o para obtener información general de la comunidad de usuarios.

El tratamiento de datos se ha convertido en un nuevo motor económico y un nuevo modelo de negocio, algunas de sus aplicaciones más relevantes son:

- Big-data
- Minería de opinión/reputación
- Análisis de tendencias
- Sistemas de dominios específicos
    - Información biomédica
    - E-health
    - Viajes

Imaginemos una empresa que actúa como intermediaria entre restaurantes que ofrecen comida a domicilio y los usuarios finales. Esta empresa provee de un servicio de repositorio de un conjunto de restaurantes. La información que generan los usuarios, como la localización desde la que se está usando el servicio o el tipo de comida que consume, pueden ser usada para mejorar su experiencia e individualizarla. En función de su localización, se mostrarán restaurante que entreguen pedidos en esa zona geográfica. Además, según sus datos de consumo se le podrán realizar recomendaciones que se ajusten gustos.

Pero esos mismos datos, procesados sobre el conjunto de la comunidad de usuarios del servicio, pueden generar valor añadido. Un análisis de

tendencias de consumo por zona geográfica es información que puede ser de gran utilidad en ámbitos como el marketing, publicidad, sociología o en consultoría de negocio.

En el mundo actual, la ciencia de datos y la gestión de la información se postulan como una de las grandes áreas de conocimiento, cuyo desarrollo e investigación ayudará a definir la sociedad del futuro.

1.3. Motivaciones

Una gran parte de la información digital está expresada en lenguaje natural. Las páginas web, blogs o redes sociales como Twitter son claros ejemplos de información en formato texto. Poder gestionar información expresada en lenguaje natural es uno de los grandes retos a los que se enfrentan las ciencias de la computación. La necesidad de contar con herramientas computacionales que permitan entender el lenguaje humano, ha dado lugar a un área de conocimiento conocido como Procesado de Lenguaje Natural(PLN).

El procesamiento del lenguaje natural, es un área que incluye disciplinas y herramientas de la ciencia computacional, como la inferencia estadística o el aprendizaje automático, junto con la lingüística aplicada. El principal objetivo de el PLN es la comprensión y el procesamiento por ordenador de información expresada en lenguaje humano[5].

Cuando nos enfrentamos a grandes colecciones de documentos de los que no se posee información alguna, los modelos de tópicos son una de las herramientas más útiles del PLN.

El Modelado de Tópicos es un conjunto de herramientas matemáticas que permiten extraer información y categorizar grandes conjuntos de documentos. El modelado de tópicos tiene como objetivo hallar a través de algoritmos estadísticos, los principales temas o tópicos de colecciones de documentos[6].

Dentro de los modelos de tópicos, el Latent Dirichelt Allocation(LDA)[7], es uno de los más empleados en la actualidad. El modelo LDA es un modelo probabilístico generativo de tópicos. Los documentos de un corpus se representan como una combinación aleatoria sobre los tópicos latentes, donde cada tópico está caracterizado por una distribución de probabilidades sobre un conjunto fijo de palabras[6].

Por otro lado, los Blogs presentes en Internet son una fuente de información desestructurada, sin categorizar y expresada en lenguaje natural.

La pregunta de investigación que motiva el desarrollo de este proyecto es si la representación de publicaciones de Blogs, como combinación de tópicos latentes , es susceptible de aportar información entendible a nivel humano que ayude a caracterizar tanto el contenido de los Blogs como las relaciones que se establecen entre los mismos.

1.4. Objetivos

El objetivo principal es dar una respuesta justificada a la pregunta que motiva la investigación. Sin embargo, para introducir el enfoque que se aplica a la resolución del problema, se presentan a continuación una lista de objetivos que guían a la consecución del objetivo principal:

Modelado de tópicos

Implementar un modelo de tópicos LDA aplicado a publicaciones de una colección de blogs de la plataforma Blogger.

Análisis del funcionamiento del LDA

El modelo de tópicos LDA es un método de aprendizaje no supervisado, basa su funcionamiento en un modelo generativo probabilístico. Su implementación depende de una serie de parámetros libres que influyen en el modelo resultante. Se pretende analizar los efectos de estos parámetros en la generación de modelos sobre la muestra. Obtener un modelo de tópicos óptimo, para la muestra bajo estudio, es fundamental si se pretende analizar la relevancia de la información a la salida del mismo.

Visualización de tópicos

Chaney yBlei[11] recalcan la importancia de la visualización de resultados en el modelado de tópicos. Los modelos de tópicos son herramientas estadísticas de alto nivel, el usuario debe escrutar distribuciones numéricas para interpretar y explorar los resultados. Es aquí donde nos planteamos establecer una visualización que permita interpretar la composición y relación entre tópicos a la salida del modelo

Visualización de documentos

De la misma forma que ocurre con la composición y relación entre tópicos, se pretende visualizar el conjunto de documentos de la muestra. Atendiendo a medidas de distancia entre documentos se plantea una visualización vía Escalamiento Multidimensional (Multidimensional Scaling, MDS)1.

Visualización de grafos y detección de comunidades

Aplicar métricas de similitud entre documentos para construir matrices pesadas dispersas de adyacencia. Implementar medidas de centralidad de red, presentar una visualización en grafo de las relaciones entre documentos y aplicar algoritmos de detección de comunidades sobre el grafo resultado.

Estos objetivos conforman el esqueleto de la investigación. El análisis de resultados en dichos objetivos, fundamentará la justificación de la respuesta a la pregunta de investigación.

[1] El MDS es una técnica multivariante de interdependencia que trata de representar en un espacio geométrico de pocas dimensiones las proximidades entre un conjunto de ítems, en nuestro caso documentos de texto.

1.5. Marco Regulatorio

Sobre el marco regulatorio aplicable a la presente investigación cabe mencionar los siguientes aspectos:

El derecho a la protección de datos personales constituye es un derecho fundamental. Está recogido en la Ley Orgánica de Protección de Datos.

En este trabajo se descargan contenidos de blogs de la plataforma Blogger para constituir la muestra bajo estudio. La plataforma Blogger es propiedad de Google.

Google cuenta con una API propietaria que permite la descarga de contenidos, asegurando la protección de datos personales de los creadores.

Por otro lado, cabe valorar el uso de esos contenidos, desde un punto de vista de propiedad intelectual. Dado que no se pretende construir ningún modelo para explotación comercial y simplemente se emplean los contenidos descargados para fines investigativos, se puede asegurar que no se incurren ilegalidades en este sentido.

Capítulo 2

2. Estado del Arte

El objetivo de este capítulo es introducir al lector en las diferentes áreas de conocimiento que confluyen en la presente investigación, además de proporcionar los conocimientos necesarios para poder entender el desarrollo y las conclusiones de este trabajo.

Para ayudar a un mejor entendimiento del problema planteado, se irán introduciendo de manera ordenada las diferentes áreas de conocimiento y tecnologías que quedan integradas en la implementación. El planteamiento del problema, estructurado en bloques encadenados, será el hilo conductor del capítulo.

Por último, para concluir el capítulo, se presentan una serie de investigaciones afines que proponen soluciones a problemas similares al planteado en este proyecto de investigación.

2.1. Introducción

Atendiendo a la pregunta de investigación planteada, cabe preguntarse cuál es proceso y jerarquía de la información.

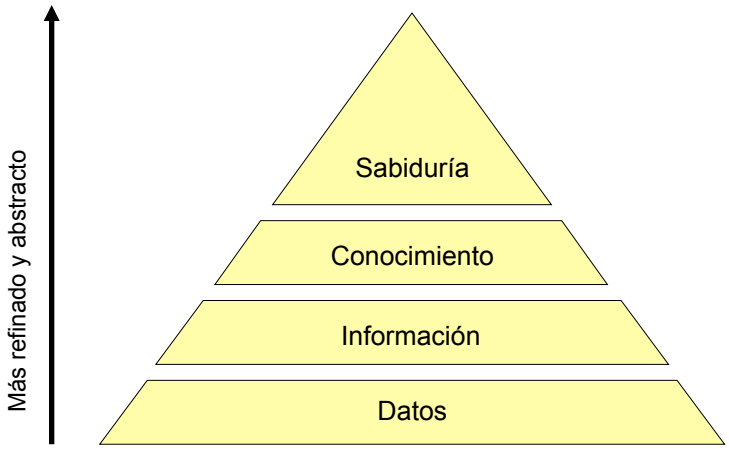


Figura 1

- *Datos*: El componente sin procesar de la información.
- *Información*: Datos organizados y presentados de una manera concreta.
- *Conocimiento*: Información sobre la que se puede actuar, es decir, creencias verdaderas justificadas.
- *Sabiduría*: Conocimiento procesado e integrado, representando comprensión a "alto nivel".

En el problema planteado, las publicaciones de los Blogs evaluados conformarán los datos. Por tanto, como ya se ha introducido, encontramos documentos de texto expresados en lenguaje natural como datos a la entrada del modelo.

El modelado de tópicos LDA será el sistema empleado para caracterizar y organizar los datos. La capacidad de este modelo para presentar datos, que aporten información útil para el entendimiento y caracterización de una colección de publicaciones de Blogs, es lo que se pretende evaluar.

2.2. Blogger

Atendiendo a la definición de la RAE, un Blog es un sitio web que incluye, a modo de diario personal de su autor o autores, contenidos de su interés, actualizados con frecuencia y a menudo comentados por los lectores.

En un Blog normalmente el contenido está estructurado en publicaciones o entradas que se suceden, ordenadas cronológicamente según su fecha de publicación. Por lo tanto, el contenido de un blog es susceptible de ser heterogéneo en cuanto a su temática.

Este hecho nos hace suponer que los modelos de tópicos no permitirán hacer una idónea categorización temática del conjunto de evaluación. Sin embargo, lo que se pretende estudiar es si la composición y relación entre los tópicos generados, junto con la relación entre documentos representados en un espacio de tópicos permite caracterizar y perfilar una colección de Blogs.

Para la composición de la muestra de este trabajo se emplearán Blogs de la plataforma Blogger. Esta plataforma, propiedad de Google, permite crear y publicar en una página Web con formato Blog. Mediante el uso de esta plataforma, el autor y gestor del contenido no requiere de conocimientos en tecnologías web. Blogger es una plataforma de alto nivel que permite publicar contenido sin necesidad de escribir código o instalar programas de servidor. Los blogs de Blogger quedan alojado en servidores de Google bajo el dominio *blogspot.com*.

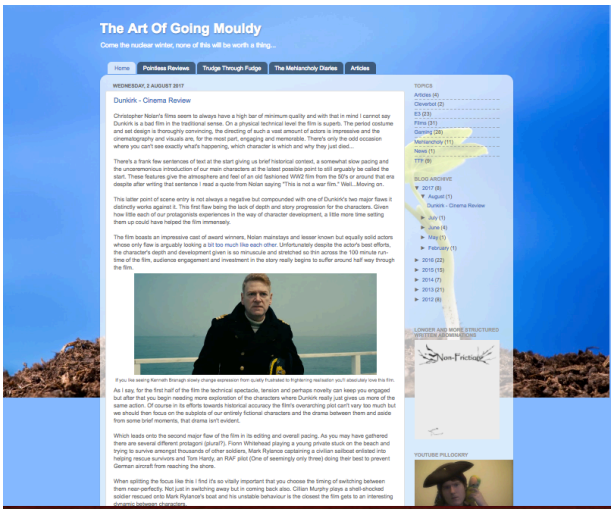


Ilustración 1. Blog de la plataforma Blogger, fuente: http://mouldywriting.blogspot.com.es/?expref=next-blog

2.3. Web Crawling y Web Scraping

Como primer punto de interés cabe preguntarse que tecnologías permiten la extracción de texto de los diferentes Blogs presentes en Internet.

Web Crawling

Un Web crawler, también conocido como araña web, es un programa informático que efectúa inspecciones de la Worl Wide Web de manera repetitiva. Un crawler es un tipo específico de programa bot en el cual un script capta y almacena información de una web de forma automática y mucho más rápido de lo que lo haría una persona.

Un crawler comienza visitando una serie predeterminada de URLs, reconoce los hipervínculos presentes en dichas webs y los añade a la lista de URLs a visitar de manera iterativa. Por lo tanto, un crawler es capaz de inspeccionar la World Wide Web descargando un conjunto de páginas web de manera recurrente.

Por tanto el uso de crawlers permite recolectar un conjunto de webs de manera sencilla y automática. Para la obtención del conjunto de Blogs que constituirán la muestra de la investigación se implementa un sencillo crawler, en el siguiente capítulo se explica con mayor profundidad el procedimiento seguido.

Web Scraping

En este punto, tendríamos un conjunto de direcciones URL que se corresponderían con los Blogs de la muestra. Ahora bien, necesitamos extraer los documentos de textos que componen las publicaciones de los distintos Blogs.

Web scraping es una técnica software para navegar por documentos en formato HTML. Permite poder obtener la información pretendida dentro del conjunto de datos que presenta una web. Mediante las técnicas de scraping somos capaces de identificar los contenidos de una página web para su selección y posterior procesado.

En el siguiente capítulo se describe la implementación que permite obtener las publicaciones, de un Blog a partir de su dirección URL.

2.4. Modelado de tópicos LDA

Sin entrar a explicar de una forma rigurosa el funcionamiento del modelo Latent Dirichelt Allocation, pues no es objeto de este trabajo, será necesario introducir su bases de funcionamiento. El modelo depende de una serie de parámetros:

- Hiperparámetros α y β de la distribución multivariada de Dirichlet
- Número de tópicos

La construcción de un modelo de tópicos mediante cualquier algoritmo que implemente LDA está supeditada a que estos parámetros estén predeterminados. Analizar el funcionamiento del LDA, para una colección de Blogs, implica comprender el efecto de estos parámetros sobre el modelo resultado.

Latent Dirichlet Allocation

LDA es un modelo probabilístico generativo de tópicos. Los documentos de un corpus se representan como una combinación aleatoria sobre tópicos latentes, donde cada tópico está caracterizado por una distribución de probabilidades sobre un conjunto fijo de palabras[6].

Algunas definiciones relevantes:

- *Corpus*: Colección de documentos a modelar
- *Palabra*: Unidad básica de información
- *Vocabulario*: Conjunto discreto de las palabra que componen un corpus
- *Documento*: Texto expresado en lenguaje natural
- *Frecuencia de término*: Representa el número de veces que se repite una determinada palabra en un documento concreto, se denota como Tf.

Representación de datos a la entrada del modelo:

Dentro del PNL, el modelado de tópicos basa su funcionamiento en la representación de documentos mediante Modelo de Espacio Vectorial(VSM). Salton et al.[8] presentaron este modelo de representación en 1975 y es uno de los más usados en la actualidad. Una buena interpretación de este modelo es:

"En VSM cada documento se identifica como un vector de rasgos en un espacio en el cual cada dimensión corresponde a términos indexados distintos (palabras). Un vector documento dado, en cada componente tiene un valor numérico para indicar su importancia" [9]

VSM utiliza un enfoque léxico, es decir, se analizan las palabras de un documento individualmente, no se establece relaciones semánticas entre palabras. Por lo tanto, tampoco importa su orden de aparición en el documento.

Como primer paso, se determina el conjunto de palabras que componen el vocabulario[2]. Cada documento de texto queda definido como una secuencia de palabras. Esto es lo que se conoce como Bolsa de Palabras. Cada documento es una bolsa de palabras, es decir, un subconjunto del vocabulario . Por ejemplo:

"esto es un ejemplo de prueba expresado en lenguaje natural"

d =[esto , es , un , ejemplo , de , prueba , expresado , en , lenguaje , natural]

Ahora bien, dado un vocabulario V de tamaño N, se representa cada documento como un vector de tamaño N, donde en cada posición aparece el Tf de esa palabra para ese documento. El valor numérico que se emplea para indicar la importancia de la palabra es el Tf.

El corpus queda representado como una matriz de tamaño DxN, donde D es el número de documentos del corpus. Cada fila de esta matriz es un documento representado en VSM.

Para poder explicar de manera sencilla el funcionamiento del LDA, se recurre a la explicación que plantea Moya García[10].

[2] En el siguiente capítulo se explica en detalle como se construye el vocabulario en el modelo implementado.

Notación:

- K: Número de tópicos.
- N: Tamaño del Vocabulario.
- M: Número de documentos .
- α: Parametro de Dirichlet:
  Vector de dimensión K que describe el conocimiento a priori que se tiene sobre como los temas se distribuyen en los documentos.
- β: Parámetro de Dirichlet:
  Vector de dimensión K que describe el conocimiento a priori que se tiene sobre como las palabras se distribuyen en cada tema.
- θ: Distribución de probabilidad de que un documento pertenezca a un tema.
- Z: Distribución de probabilidad de que una palabra pertenezca a un tema.
- W: Identifica todas las palabras en todos los documentos.
- φ: Distribución de probabilidad de que dado un tópico salga una palabra.

LDA proporciona dos matrices de distribuciones de probabilidad $P(w \mid z)$ y $P(z \mid \theta)$, la primera distribuciones es la de que dado un tema salga una palabra y la segunda es la de que un documento pertenezca a un tema. En LDA cada documento se representa por un vector que sigue una distribución multivariada de Dirichlet. A continuación se presenta una figura explicativa que propone Moya García[10].

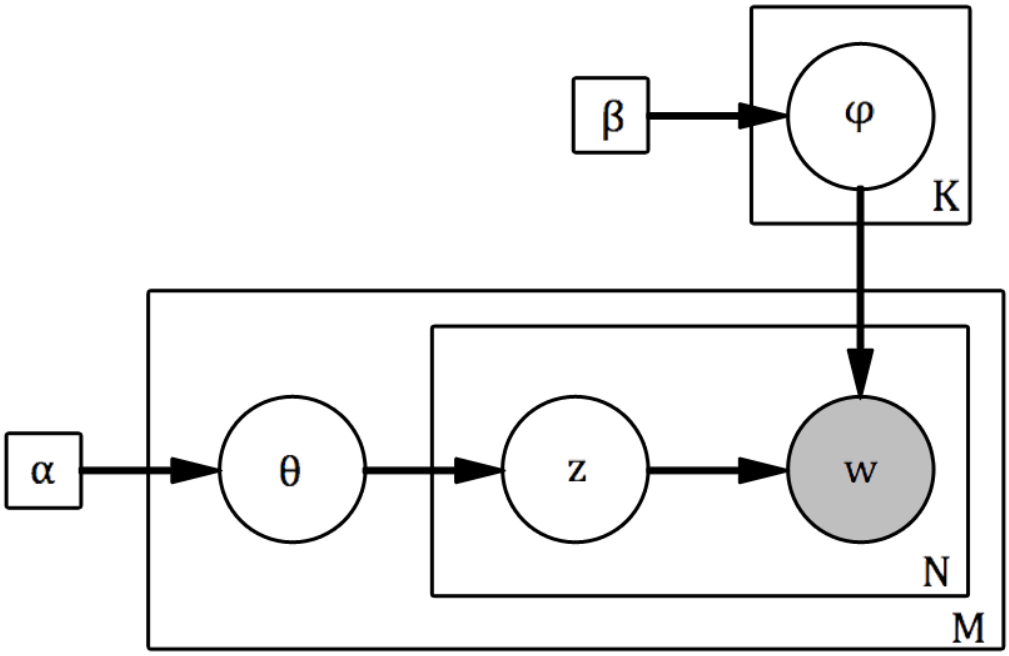


Figura 2: Diagrama de bloques LDA

Atendiendo a lo que expone Blei et al.[7] los parámetros α y β de Dirichlet, suelen ser parámetros definidos a priori que pueden tomar valores dentro del intervalo abierto (0,1). Si se presupone que el documento está categorizado[3], se recomienda valores de α pequeños. Por el contrario si se presupone que un documento va a tratar sobre varios temas, se recomienda poner un valor de α grande. El valor β depende de la cantidad de palabras que definen un tópico. Se recomienda que si un tópico va necesitar de pocas palabras el valor de β sea muy pequeño. Por el contrario, si un tópico requiere muchas palabras, se recomienda que β tome un valor grande.

[3] Con categorizado nos referimos a que trate sobre un tema bien definido. Por ejemplo, documentos sobre economía.

Representación de datos a la salida del modelo

La implementación de un modelo LDA transforma los datos a la entrada para establecer la siguiente representación:

> Tópicos: LDA devuelve una matriz de tamaño Kx|V|, es decir, se definen cada tópico k_i :{i $\in$ (0,K-1)} como una composición del aporte de cada palabra del vocabulario V en el tópico.
>
> Documentos: LDA devuelve una matriz de tamaño DxK, es decir, se define cada documento d_i :{i $\in$ (0,D-1)} como una composición del aporte de cada tópico del conjunto K en el documento.

Visualizar Modelos de Tópicos

El objetivo de presentar una visualización de tópicos, es representar los datos a la salida del modelo de forma interpretable. Como establecen Chaney yBlei[11] los objetivos de la visualización deben ser:

- Resumir el corpus visualizando la composición de los tópicos.
- Revelar las relaciones entre los documentos evaluados y los tópicos descubiertos.
- Revelar relaciones entre los documentos evaluados.

Estos objetivos establecen las bases de la extracción de información de una colección de documentos empleando un modelo de tópicos.

En la literatura que versa sobre visualización de tópicos se suele hacer especial hincapié en el primero de los objetivos presentados. Es decir, las visualizaciones suelen centrarse en la de la composición de los tópicos descubiertos. Este tipo de visualizaciones permiten comprender el corpus de manera global, pero no prestan atención a los documentos a nivel individual. Por tanto, se desatienden las relaciones que se establecen entre documentos.

En este trabajo se pretende cubrir los tres objetivos para intentar responder a la pregunta que motiva la investigación:

¿Es el modelo LDA susceptible de aportar información, entendible a nivel humano, que ayude a caracterizar tanto el contenido de los Blogs como las relaciones que se establecen entre los mismos?

Sivert y Shirley[12] proponen un método –*LDAvis*- para la visualización de la composición de los tópicos. En este trabajo se emplea este método de visualización, se introduce a continuación sus bases de funcionamiento de acuerdo a lo expuesto por Sivert y Shirley[12].

La herramienta *LDAvis* permite una visualización interactiva que ofrece las siguientes posibilidades:

- Tópicos representados mediante círculos en el plano de dos dimensiones, donde el área de cada circulo representa la predominancia de cada tópico en el corpus. Los tópicos son ordenado en orden decreciente respecto a su predominancia.

- La posición de cada tópico se computa mediante la distancia entre tópicos, usando MDS para proyectar estas distancias al plano de dos dimensiones.

- Se representa mediante diagramas de barras las palabras más relevantes para el tópico seleccionado. Se superponen una barra roja y otra azul para cada término que representan:

    - Azul: Tf de la palabra evaluado en todo el corpus, es decir, el número de veces que se repite esa palabra en el conjunto de documentos evaluados.
    - Rojo: Tf estimada de la palabra en el tópico seleccionado.

A continuación se presentan tres figuras que ilustran la apariencia que presenta la herramienta *LDAvis.*

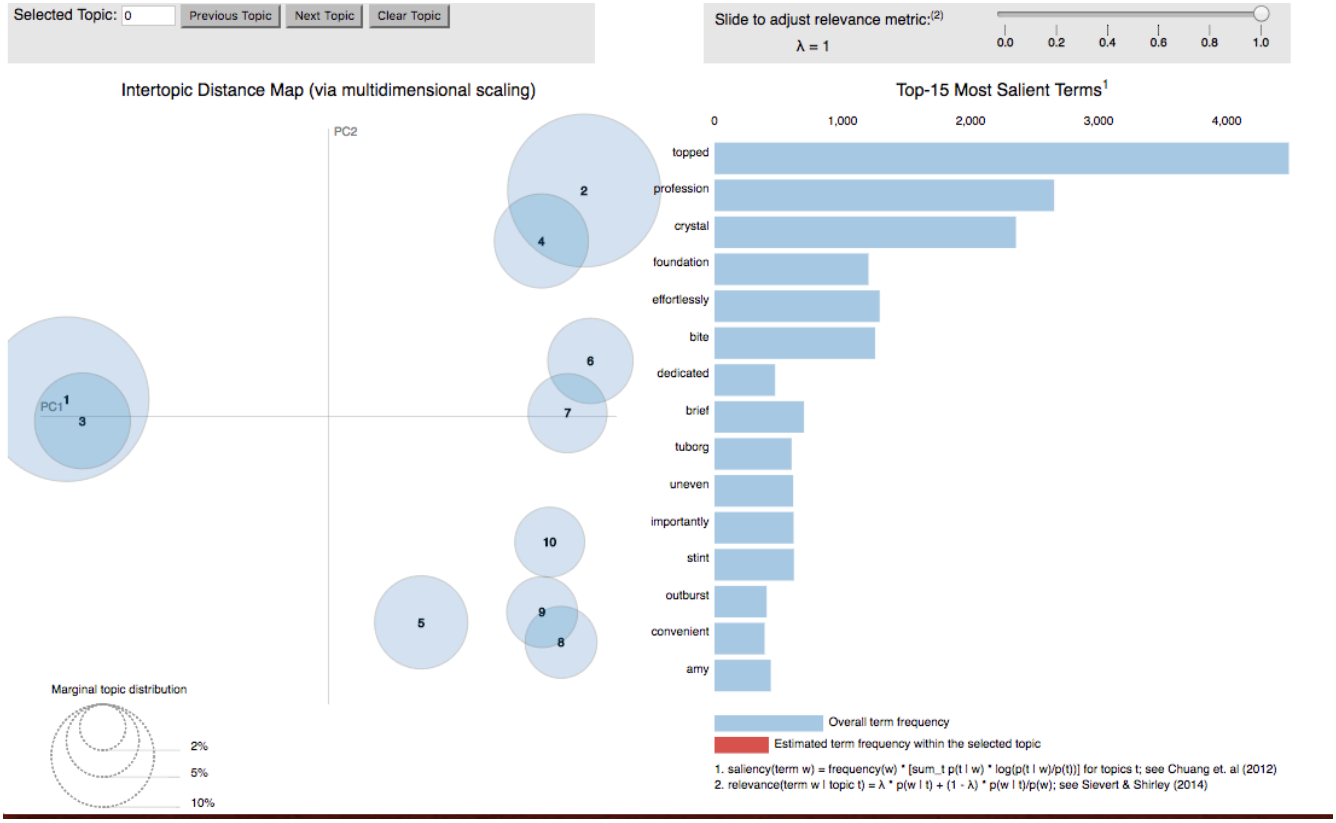


Figura 3: Visualización de tópicos vía *PyLDAvis*

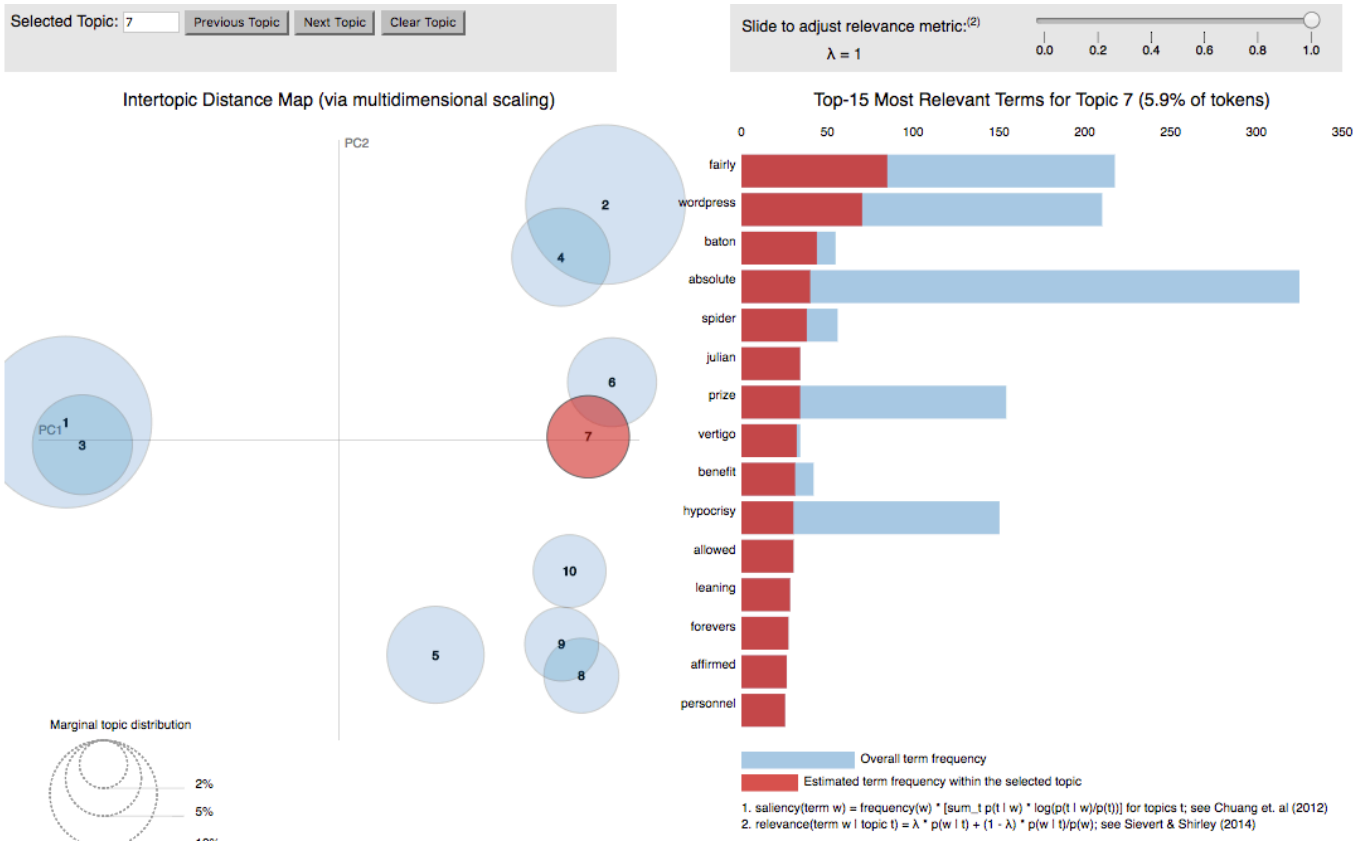


Figura 4: Visualización de composición de tópico vía *PyLDAvis*

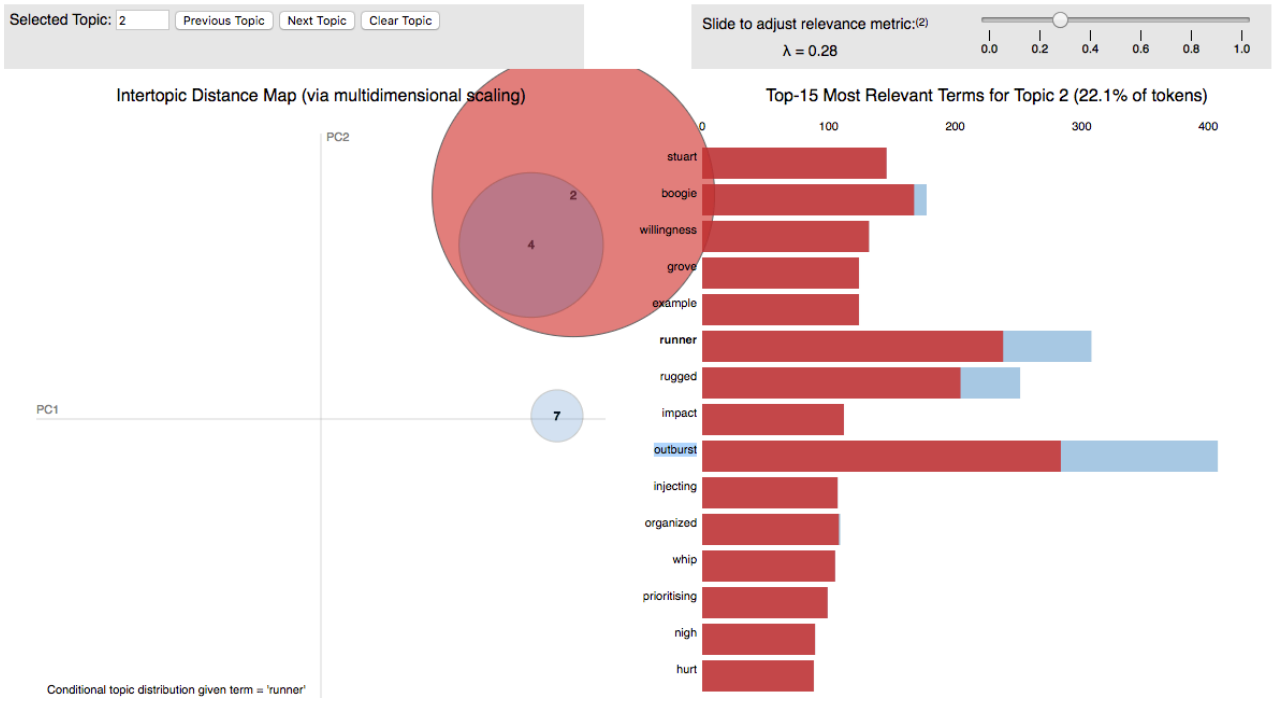


Figura 5: Pertenecia de palabra la outburst a los distintos tópicos

El diagrama de barras de la figura 3, en la que no se ha seleccionado ningún tópico, representa las 15 palabras con mayor prominencia. La medida de prominencia se corresponde con la propuesta por Chuang et al.[14]. Se calcula de acuerdo a:

$$Prominencia(w) = Tf_{corpus(w)} * \sum_{i=0}^{K-1} p(k_i|w) * \log\left(\frac{p(k_i|w)}{p(k_i)}\right)$$

Como se observa en la figura 4, cuando se selecciona un tópico se presentan sobre el diagrama de barras las medidas antes explicadas.

La figura 5 ilustra que si seleccionamos una palabra concreta, podemos ver gráficamente a que tópicos pertenece. El área de los círculos, pasa a indicar el grado de pertenencia de la palabra al tópico.

Por último cabe mencionar, que como se observa en la parte superior izquierda, en la herramienta de visualización LDAvis hay un parámetro libre $\lambda \in (0,1)$. Este parámetro influye a la hora de ordenar los términos más relevantes de cada tópico. Este parámetro se extrae de un aporte de Sivert y Shirley[12], que propone una medida de la relevancia de una palabra en un tópico *k* de acuerdo a:

$$Relevancia(w|k) = \lambda * \log[p(w|k)] + (1-\lambda) * \log[\frac{p(w|k)}{p(w)}]$$

Como se puede observar el parámetro λ determina el peso dado a la probabilidad del término w bajo el tópico k respecto a su lift[4].

Si λ=1 la relevancia queda definida por la probabilidad de la palabra w dado el tópico k. Sivert y Shirley[12] determina un λ=1/3 como el óptimo para el caso de estudio que presentan.

Evaluación e interpretación de Modelos de Tópicos

En la evaluación de los modelos de tópicos se suele asumir que el espacio latente de tópicos es semánticamente significativo. Bajo este supuesto se pueden implementar medidas de evaluación del modelo basadas en:

Perplejidad del modelo probabilístico: Cuando se pretende ajustar un modelo probabilístico a un conjunto de datos, se puede evaluar el modelo mediante validación cruzada. Se emplean los datos de validación, los que no se usan para entrenar, para calcular la perplejidad sobre el modelo propuesto. Dado un modelo de una distribución de probabilidad desconocida p. El modelo q, propuesto sobre muestras de entrenamiento de la distribución desconocida p, puede evaluarse mediante la perplejidad para un conjunto de muestras de validación $x_1, x_2, \dots, x_N$. La perplejidad se define como:

$$H(q) = 2^{-\sum_{i=0}^{N-1}\frac{1}{N}\log_2 q(x_1)}$$

Valores menores de perplejidad indican que el modelo define mejor las muestras de validación.

Held-out Likelihood: Si se lleva a cabo un proceso de validación cruzada *Held-out Likelihood* es la verosimilitud del modelo, evaluada sobre los parámetros estimados, empleando los datos de validación.

[4] *Lift* es una medida de interés que mide hasta qué punto ocurren conjuntamente w y k, más o menos de lo esperado, si fuesen independientes.

Sin embargo como establece Chang et al.[15] estas medidas no evalúan lo significativo semánticamente que es el modelo. Esto es importante de entender. Estas medidas evalúan lo bueno que es el modelo predictivo probabilístico, pero no abordan objetivos de evaluación humana.

El Modelado de Tópicos debe evaluarse también desde la perspectiva del entendimiento humano. Los tópicos deben poder expresar información entendible a nivel humano. Esto conlleva un análisis cualitativo y cuantitativo del modelo.

En cuanto a lo relativo a la presente investigación, se requiere descubrir cuál es el número de tópicos óptimos que define una colección. En el problema propuesto, los documentos a evaluar son publicaciones de blogs. La heterogeneidad de este tipo de documentos en cuanto a su temática hace que sea complejo evaluar cualitativamente el número óptimo de tópicos.

Por este motivo, se emplearán valoraciones cualitativas, observando la composición de los tópicos con *LDAvis*. Pero a su vez se empleará una medida cuantitativa que permite descubrir el número óptimo de tópicos que define un corpus. Por tanto en esta parte se hará una análisis cualitativo y cuantitativo conjuntamente.

Arun et al.[17] propone una medida para descubrir el número óptimo de tópicos en un determinado corpus. La idea es evaluar un conjunto de modelos con distinto número de tópicos predefinidos, para encontrar el número que mejor caracteriza el corpus evaluado. Esta medida será la empleada para evaluar el número óptimo de tópicos.

La medida propuesta por Arun et al.[17] basa su funcionamiento en el siguiente hecho:

Si las matrices que devuelve LDA a su salida no fuesen estocásticas tendríamos:

- Matriz M1 de K tópicos como número de filas y tantas columnas como palabras en el vocabulario V. El elemento (i,j) indicaría el número de veces que la palabra j ha sido asignada al tópico i.
- Matriz M2 de D documentos como número de filas y K tópicos como número de columnas. El elemento (i,j) indicaría el número de veces que una palabra del tópico j ha sido asignado a alguna palabra del documento i.

Atendiendo a este hecho es intuitivo comprobar que:

$$\sum_{w=0}^{W-1} M_1(k,w) = \sum_{d=0}^{D-1} M_2(d,k)$$

Esto no es más que el número de palabras asignadas a cada tópico visto de dos formas diferentes. Sin embargo, cuando se normalizan estas matrices por filas, como hace LDA, esta igualdad deja de ser válida.

La idea detrás de lo propuesto por Arun et al.[17] es aprovechar el hecho de que las dos sumas representan la proporción de tópicos asignados al corpus y, por tanto, pueden compararse entre sí.

Sin embargo, una mera comparación entre estos valores no es independiente del número de tópicos considerados en el modelo. Por lo tanto, se busca una medida, que al tratar de comparar las propiedades similares de estas matrices, sea baja sólo cuando se alcance el número de óptimo de tópicos.

La medida que se propone es:

$$Medida\ Propuesta = D_{KL}(C_{M1}||C_{M2}) + D_{KL}(C_{M2}||C_{M1})$$

Donde:

D_{KL} representa la divergencia de Kullback-Leibler[5]

C_{M1} son los valores singulares de la matriz M1, por lo tanto, es un vector de K componentes.

C_{M2} es el resultado de multiplicar L*M2 y normalizar el vector resultado, donde L es un vector de dimensiones 1xD. Los valores contenidos en L representan el número de palabras diferentes que aparecen en cada documento. De esta forma C_{M2} es un vector de K componentes.

Evaluando esta medida para un conjunto de tópicos predefinidos, el valor óptimo será para el cual la distancia entre los vectores $C_{M1}\ y\ C_{M2}$ sea mínima.

[5] Como se explica en el siguiente apartado la divergencia de Kullback-Leibler es una medida que se puede interpretar como la distancia entre dos distribuciones de probabiblidad.

2.5. Métricas de Similitud

Si recuperamos los objetivos expuestos de una visualización de tópicos:

- Resumir el corpus visualizando la composición de los tópicos.
- Revelar las relaciones entre los documentos evaluados y los tópicos descubiertos.
- Revelar relaciones entre los documentos evaluados.

Hasta ahora hemos abordado el primero de los puntos. Para evaluar los otros dos, necesitamos establecer métricas de similitud entre documentos. Si recordamos, a la salida de los modelos LDA, los documentos quedan representados como vectores en los que cada componente indica la pertenencia del documento a un tópico concreto. Es necesario, basándonos en esta representación, establecer lo que se parece cada documento a todos los demás.

Las medidas de similitud establecen el parecido entre dos vectores, en nuestro caso vectores de una distribución de Dirichlet.

- Si $x=y \rightarrow$ similitud$(x,y)=1$
- Si $x \neq y \rightarrow 0 \leq$ similitud$(x,y) \leq 1$

La similitud también puede interpretarse como:

- distancia$(x,y)=1-$similitud(x,y)

La distancia *d,* se define formalmente como una función matemática que verifique:

- No negatividad: $d(x,y) \geq 0$
- Simetría: $d(x,y)=d(y,x)$
- Desigualdad triangular: $d(x,z) \leq d(x,y)+d(y,z)$
- $d(x,x)=0$
- $d(x,y)=0 \rightarrow x=y$

Una medida de distancia muy empleada y intuitiva de entender es la distancia euclídea. Pero para nuestro caso particular, como bien explica Moya García[10] , la métrica euclídea no ofrece buenos resultados. Esto se debe a que los vectores bajo estudio son distribuciones de probabilidad multivariada.

Para valorar la distancia entre dos funciones de distribución de probabilidad se cuenta con la divergencia de Kullback-Leibler, aunque no se considera como una distancia ya que no es simétrica, ni cumple la desigualad triangular. En la presente investigación se emplea la versión simétrica de la divergencia de Kullback-Leibler como medida de distancia.

$$dist_{KL}(X||Y) = divergencia_{KL}(X,Y) + divergencia_{KL}(Y,X)$$

Donde la divergencia de Kullback-Leibler, para dos distribuciones de probabilidad *X* e *Y* de una variable aleatorio discreta, se define como:

$$D_{KL}(X||Y) = \sum_i X(i) * ln\frac{X(i)}{Y(i)}$$

Además como medida de similitud se emplea la divergencia de Jensen-Shanon, que está basada en la divergencia Kullback-Leibler, pero ofrece información de parecido en vez de distancia.

$$\text{JSD(X||Y)} = \frac{1}{2} * [divergencia_{KL}(X,M) + divergencia_{KL}(Y,M)].$$

Donde: $M = \frac{1}{2}(X+Y)$

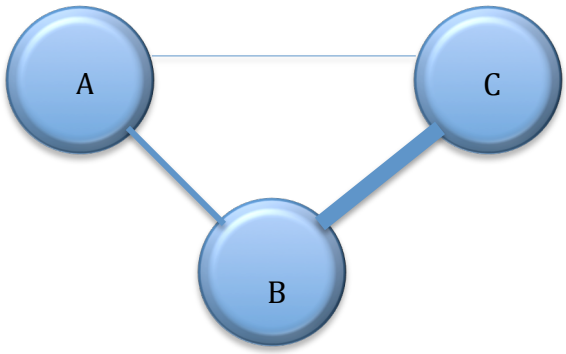
2.6. Grafos y análisis de redes

A partir de la métricas de similitud podemos construir matrices pesadas de adyacencia. Una matriz de adyacencia representa relaciones ítem a ítem de un conjunto de datos. Por tanto es una matriz cuadrada, simétrica, con unos en su diagonal. Se expone a continuación un ejemplo:

	Documento A	Documento B	Documento C
Documento A	Sim(A,A)=1	Sim(A,B)=0.5	Sim(A,C)=0.2
Documento B	Sim(B,A)=0.5	Sim(B,B)=1	Sim(B,C)=0.8
Documento C	Sim(C,A)=0.2	Sim(C,B)=0.8	Sim(C,C)=1

Tabla 1

Estas matrices son el punto de partida para la construcción de grafos y análisis de red.
Un grafo es un conjunto de nodos que quedan conectados por un conjunto de aristas. Observando la Tabla 1, los nodos serían los documentos y las aristas quedarían representadas por la métrica de similitud. Es decir las aristas representan mediante pesos la conexión o relación entre nodos.



Los grafos son un instrumento matemático que tiene una representación visual muy intuitiva. La presentación de resultados en una visualización de grafo permitirá establecer relaciones entre documentos, aportando además nodos etiquetados por su blog de procedencia y tópico predominante. Esta presentación visual pretende aportar información entendible a nivel humano que ayude a caracterizar el contenido y relaciones en una colección de blogs. Cuando hablamos de grafos compuestos por gran cantidad de nodos nos referimos a éstos como redes.

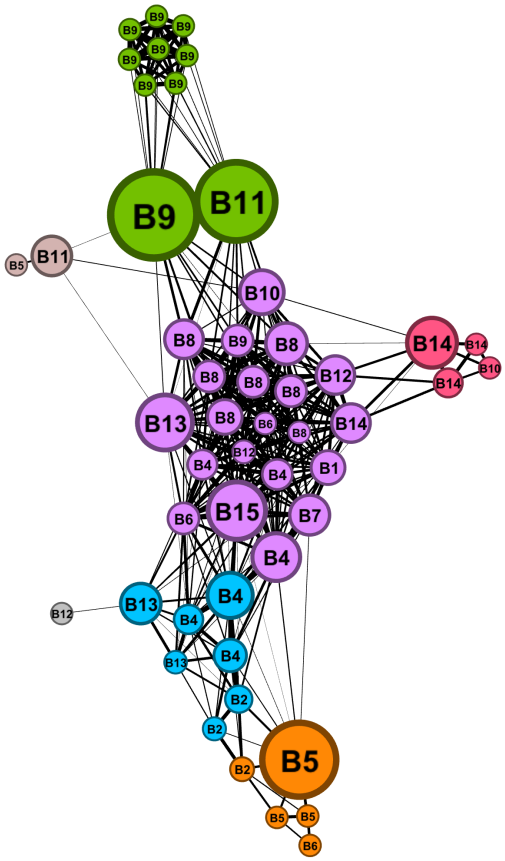


Figura 6. Red de documentos. Color de nodos representa tópico predominante. Etiqueta de los nodos indica e blog de procedencia del documento. Fuente: Resultado de esta investigación generado con el software Gephi.

Análisis de redes y detección de comunidades:

El análisis de redes se encarga de analizar redes mediante teoría de grafos. El conjunto de documentos evaluado constituirá una red de información sobre la que se pretende:

- Evaluar medidas de centralidad que ayuden a la interpretación en la visualización de la red.
- Detectar comunidades para analizar el comportamiento de las distintas agrupaciones que componen la red.

Intermediación (*Betweenness Centrality*):

La intermediación es una medida de centralidad de red. En el ámbito de la teorías de grafos y análisis de redes la centralidad es una tipo de medida que evalúa la importancia de un nodo dentro de la red.

La intermediación mide el número de veces que un nodo aparece en el camino más corto entre otro dos. Para nuestro caso concreto nos da una idea intuitiva de que documentos contienen un contenido más genérico y están relacionado con otros tópicos y blogs. En la Figura 6 el tamaño de los nodos está ligado al valor de intermediación de cada nodo. Esta medida de centralidad nos ayudará a interpretar la información que aporta la red y a ayudar a una visualización más intuitiva.

Método de Louvain para detección de comunidades:

Las comunidades son grupos de nodos dentro de una red, más densamente conectados entre sí que con otros nodos. La modularidad es una métrica que cuantifica la calidad de una asignación de nodos a una comunidad, evaluando como de densamente conectados están los nodos en una comunidad, en comparación con cómo estarían conectados, en promedio, en una red aleatoria definida adecuadamente.

El método de Louvain propuesto por Blondel et al.[16] es un algoritmo para detectar comunidades en redes. Se basa en una heurística para maximizar la modularidad. El método consiste en la aplicación repetida de dos pasos:

- El primer paso es una asignar nodos a comunidades, de manera que se favorezca la optimización local de las modularidades.
- El segundo paso es la definición de una nueva red, en términos de las comunidades encontradas en el primer paso.

Estos dos pasos se repiten hasta que la modularidad deja de crecer en la reasignación de nodos a comunidades que se realiza del primer paso.

El método de Louvain logra modularidades comparables a los algoritmos preexistentes, normalmente en menos tiempo, por lo que permite el estudio de redes mucho más grandes. También revela generalmente una jerarquía de las comunidades en diversas escalas, y esta perspectiva jerárquica puede ser útil para entender el funcionamiento global de una red. El método de Louvain será el empleado para la detección de comunidades en el modelo implementado.

2.7. Investigaciones afines

Existe una amplia literatura sobre investigaciones, empleando algoritmos que implementen LDA, para caracterizar colecciones estructuradas. Normalmente los objetivos de estas investigaciones están más en la línea de la clasificación automática de textos mediante representación de documentos como combinación de tópicos latentes.

Sin embargo, también se han encontrado propuestas que se acercan más al enfoque que en esta investigación se plantea. Ramage et al.[22] presentan una implementación escalable y parcialmente supervisada del modelo LDA para caracterizar el contenido que se publica en la red social Twitter. Twitter es una red social en las que las publicaciones son textos de menos de 140 caracteres como máximo. Twitter representa un ejemplo de los que se conoce como microblogging.

Esta investigación emplea el método LDA-Etiquetado (Labeled-LDA, L-LDA) propuesto por Ramage et al.[23] para predefinir 4 categorías:

- *Substance:* Publicaciones sobre eventos, ideas o contenidos informativos.

- *Status*: Hace referencia a publicaciones de contenido personal del usuario. Publicaciones en las que los usuarios aportan información personal y subjetiva de sus vidas.

- *Social:* Relativo a comunicación social vía Twitter entre usuarios.

- *Style :* Indican tendencias más amplias de uso de lenguaje.

Partiendo de estas etiquetas, que representan los grandes temas de Twitter, implementan una caracterización de los contenidos descubriendo tópicos asociados a cada una de las etiquetas. El modelo que se propone procesa también metadatos de las publicaciones, como los hastags o las menciones a otros usuarios. De esta manera se consigue una aproximación a la caracterización del tipo de contenidos que se publican en Twitter.

Esta investigación confirma que los modelos de tópicos son capaces de aportar información útil antes colecciones desestructuradas. Aunque cabe mencionar que en esta investigación se emplean sujetos humanos para establecer, bases de conocimiento empírico, que ayudan a caracterizar las etiquetas predefinidas.

Ramage et al.[22] no solo propone un modelo para caracterizar los contenidos de Twitter. Se encuestan sujetos humanos, para establecer conocimiento empírico de los motivos por los que los usuarios de Twitter siguen a las distintas fuentes de información. También recogen información de los principales motivos por los que los encuestados dejan de seguir un perfil de Twitter. Aunando este conjunto de informaciones, basadas en encuestas con usuarios, con la caracterización de los contenidos de Twitter plantean un sistema de recomendación que ayude a los usuarios de Twitter a poder seguir perfiles que satisfagan sus intereses de información.

Los sistemas de recomendación basados en contenido, como el planteado en la investigación comentada, pueden suponer una línea futura en la presente investigación. En el capítulo 5 se profundizará sobre este aspecto.

Cabe mencionar también la investigación que desarrolla Perez-Tellez et al.[24] para realizar clustering sobre una colección de blogs. El objeto de esta investigación es proponer un método para agrupamiento de blogs, basado en la representación de documentos mediante tópicos latentes proporcionados por una implementación de LDA.

Lo más destacable de esta investigación es una aportación, que se explica a continuación, relativa al procesado de texto previo al empleo de algoritmos que implementan LDA.

Si recordamos el formato de los datos a la entrada de un modelo que implemente LDA. Tenemos que cada documento se representa como un conjunto de valores de Tf, para el conjunto de palabras diferentes existentes en el documento. La técnica propuesta *Self-Term Expanssion*[24], consiste en sustituir palabras de un documento por unas serie de términos correlacionados semánticamente. Básicamente, el funcionamiento de esta técnica explicado de forma sencilla, es evaluar pares de palabras correlacionadas semánticamente y que además aparecen en un mismo documento. El empleo de agrupaciones de términos mejora la caracterización semántica de los documentos y los tópicos latentes.

Cabe mencionar también que se ha encontrado de especial interés el trabajo de Macdonald y Ounis[25], que versa sobre la construcción de una muestra representativa de la población que componen todos los blogs presentes en la red. El objetivo que se pretende es conformar una muestra para experimentar en el área del procesado de lenguaje natural y la recuperación de información.

Capítulo 3

3. Implementación

En este capítulo se describe la implementación de los distintos módulos que componen el modelo que se presenta en este trabajo. Además, se dará detalle de los recursos utilizados para la construcción de cada uno de estos módulos.

A continuación se resumen los recursos principales empleados que se nombrará a lo largo del capítulo:

- Lenguaje de programación Python a través del entorno de desarrollo Anaconda. Las principales librerías empleadas son:
    - Google API Client: Blogger JSON API
    - NLTK
    - Numpy
    - Gensim
    - Bokeh
    - Networkx
    - Community
    - PyLDAvis
- Software de visualización de grafos Gephi

3.1. Muestra y Plan de Muestreo

Analizando la población bajo estudio, Blogs presentes en la red de Internet, se vuelve complejo establecer criterios para obtener una muestra representativa. Haciendo referencia al propio concepto del Blog, como ya se ha explicado, la forma y temática de un blog puede ser diversa. Cabe mencionar que se ha elegido evaluar blogs escritos en lengua inglesa .

En primer lugar se consideró establecer un muestreo aleatorio para la selección de la muestra .Los blogs de Blogger de dominio gratuito incluyen por defecto en la parte superior izquierda un enlace a otro blog de la plataforma. Esto se puede observar el la ilustración del final de la página. El direccionamiento a otro blog es aleatorio, es decir, el siguiente blog no está predefinido. Si se pincha múltiples veces en este enlace, cada vez devuelve un resultado diferente. Esta característica se aprovecha para obtener de manera aleatoria una colección de blogs.

*La Website Crawler Tool and Google Sitemap Generator*6 permite hacer crawling desde una dirección URL y devuelve URLs tanto internas como externas a la web origen. Partiendo de varias direcciones iniciales se obtuvieron 100 direcciones URL de blogs de la plataforma Blogger.

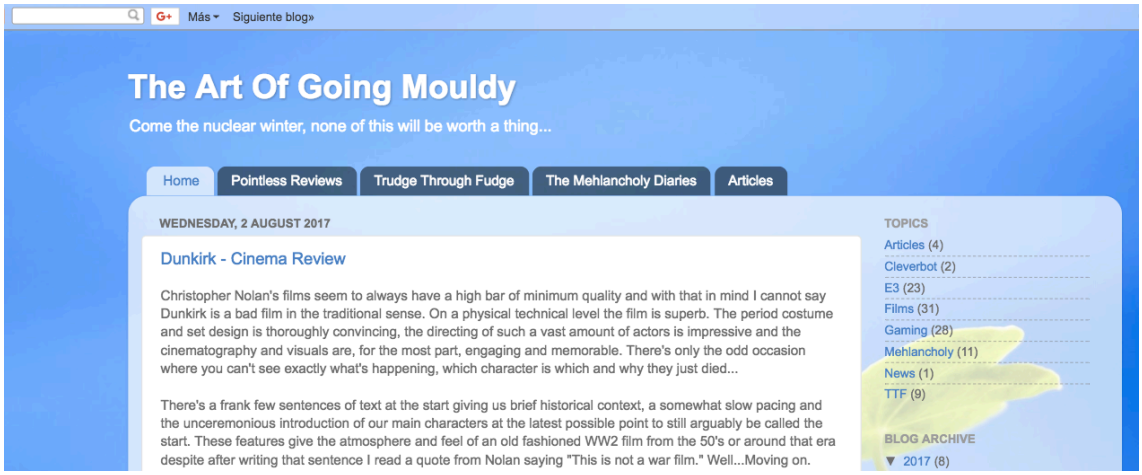


Ilustración 2. Fuente:http://mouldywriting.blogspot.com.es/?expref=next-blog

6 Fuente: http://freetools.webmasterworld.com/tools/crawler-google-sitemap-generator/#sthash.5wEzZ85u.uhUbYwY5.dpbs

Consideraciones en el muestreo:

En todo momento en este trabajo se ha identificado el contenido de un blog con información expresada en lenguaje natural. Sin embargo, esto no es siempre cierto. Los blogs admiten contenido multimedia. Está situación obligó a realizar un muestreo intencional sobre el primer conjunto de muestra.

El criterio establecido en el muestreo intencional no es otro que contar con información expresada en lenguaje natural en las publicaciones de los blogs. Los blogs evaluados pueden agruparse en tres categorías:

- Blogs con publicaciones con contenido exclusivamente multimedia:
    - Publicaciones que incluyen solo fotos o vídeo sin apenas texto.
- Blogs con publicaciones con contenido expresado exclusivamente en lenguaje natural.
- Blogs mixtos en los que las publicaciones están expresadas principalmente en lenguaje natural, pero pueden incluir contenido multimedia.

Dado el criterio expuesto para la selección de muestras, solo los dos últimos tipos de blogs cumplen. Cabe mencionar que de los blogs mixtos la información multimedia no es incluida en el modelo, se trabaja exclusivamente con el texto de las publicaciones. La evaluación temática del contenido de los blogs no es criterio de selección, para cumplir en el muestreo con las bases sentadas para el planteamiento de la investigación.

Por otro lado para seleccionar la cantidad de blogs que componen la muestra se atiende a los siguientes criterios:

- Interpretación de resultados del Modelo LDA: Ante grandes cantidades de documentos, y más si son de temáticas no muy definidas, la interpretación de resultados se vuelve compleja. Intentar caracterizar una colección grande a través de sus publicaciones puede volverse complejo. Además cabe mencionar que una de las bases para la interpretación de resultados propuesta es la visualización de grafos. Intentar interpretar relaciones visualmente en grafos con gran cantidad de nodos, cuando no se espera patrones concretos, se vuelve una tarea complicada.

- Tiempo de ejecución: Dado que se deben construir modelos con distinto número de tópicos predefinidos para su evaluación, a mayor número de documentos en la muestra más ardua y lenta se vuelve la tarea de ejecución software.

Sobre este punto de la cantidad de muestras cabe hacer una última aclaración. El propósito de la presente investigación no es generalizar los resultados al conjunto general de blogs presentes en Internet. Es decir, no se pretende evaluar la capacidad de LDA para caracterizar simultáneamente todos los blogs existentes. El objetivo principal es evaluar la capacidad de LDA de caracterizar colecciones concretas. Por este motivo la muestra debe ser suficientemente representativa, pero no se requiere una dimensión elevada de número de muestras.

Plan de muestreo:

Atendiendo a las consideraciones antes mencionada se establece el siguiente plan de muestreo:

- Muestreo aleatorio de un total de 100 blogs

- Muestreo intencional sobre un muestreo aleatorio previo:

    - Selección de dos grupos de 15 blogs, para su evaluación por separado, conformado así dos colecciones diferenciadas de blogs para su evaluación.

- Muestreo intencional de publicaciones por blog:

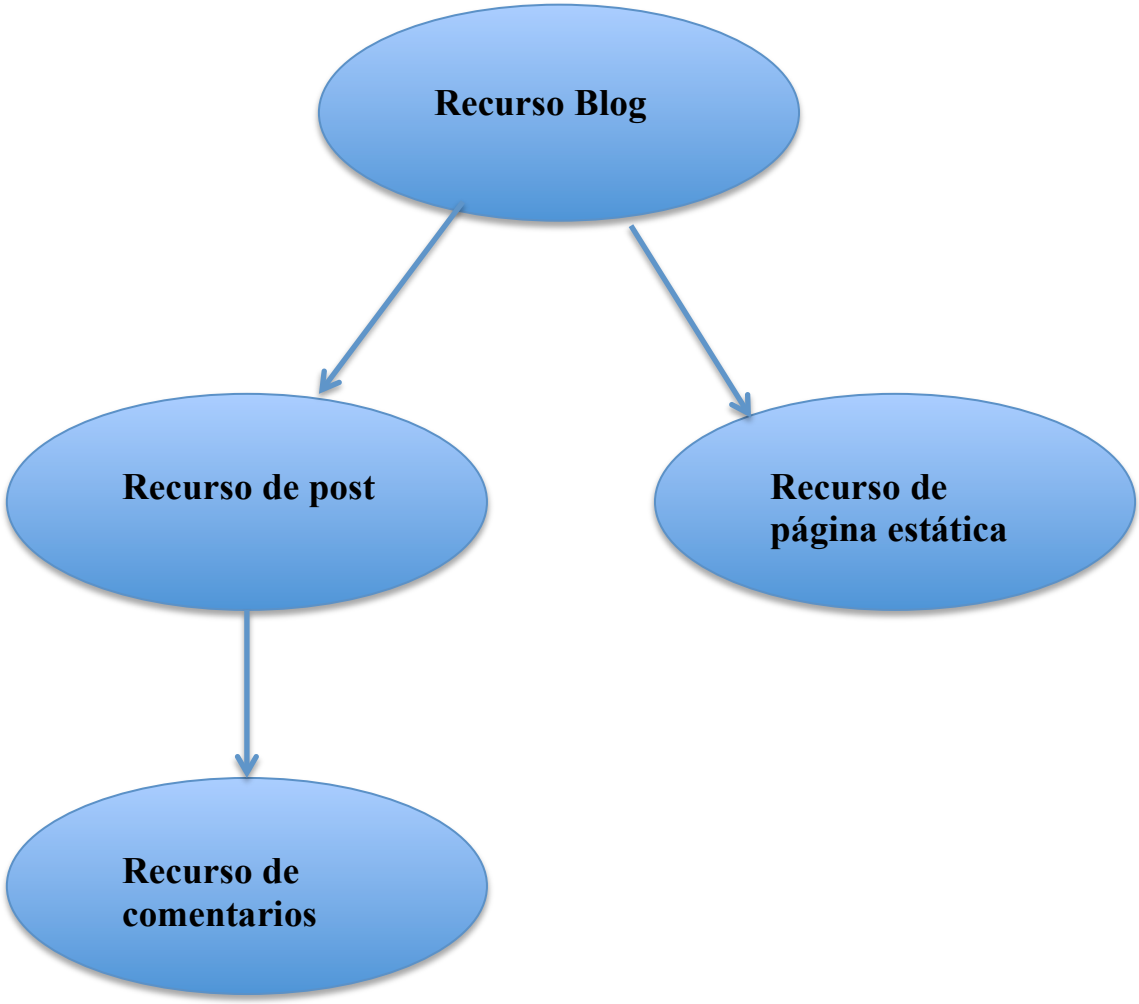
    - Criterio cronológico: Se seleccionan las 10 últimas

    - Descarga de10 publicaciones por blog para conformar los dos corpus que serán evaluados, obteniendo dos corpus de 150 documentos cada uno.

3.2. Descarga de datos

Como ya se ha mencionado, la plataforma Blogger es propiedad de Google. Google tiene una API para Python que permite trabajar con la plataforma Blogger.

Blogger JSON API permite descargar cualquier blog público. Para poder emplear la API se requiere registro en la plataforma Google API. Como su nombre indica, esta API devuelve datos en formato JSON.



El modelo de datos de esta API establece recursos, como unidades individuales de datos, con un identificador único. Los recursos principales de la API en se muestran en el diagrama anterior. En este proyecto solo se emplearán los recursos de blog y de post.

Todas las llamadas a métodos de la API se hacen mediante *Simple API Acces*. Estas llamadas no acceden a datos privados de usuario, se requiere de una API Key para poder realizar llamadas.

Para la descarga de publicaciones se estableció un script que recibía una lista de URLs de blogs y descargaba los 10 últimos post de cada blog. Se explica de manera resumido el proceso de descarga:

1. A través de la dirección URL se solicita un recurso blog.

2. Se obtiene del objeto JSON devuelto el identificador del blog.

3. Se solicita a través del identificador del blog la obtención de 10 recursos post.

4. Se obtiene de los objetos JSON devueltos el contenido de los post.

5. Almacenamiento persistente empleando ficheros para almacenar los documentos que componen el corpus.

3.3. Procesado del corpus

En este punto tenemos textos en lenguaje natural, en este apartado se explica el procesado de texto que se lleva a cabo para poder transformar los datos al formato requerido a la entrada del LDA.

La librería NLKT permite trabajar en Python con herramientas de PLN. Se emplea esta librería para el procesado del corpus. El procesado del corpus puede resumirse en las siguientes etapas:

- Tokenización
- Homogenización
- Limpieza
- Vectorización

Tokenización:

Este proceso consiste básicamente en transformar un documento e texto en un lista que contenga todas las palabras del documento separadas. Se emplea la función *word_tokenize* del paquete *nlkt.tokenize*.

Tras este proceso, cada documento queda representado por el conjunto de palabras que lo componía, pero separadas y preparadas para ser procesadas individualmente. Cada palabra es lo que se conoce como token.

Homogenización:

Si analizamos los documentos tokenizados, es decir separado en unidades textuales, verificamos que muchos tokens se corresponde con signos de puntuación que no son relevantes para el análisis semántico. El proceso de homogenización consiste en:

- Transformar los tokens para que todas las letras estén en minúscula
    - Se emplea la función *lower* de Python para representar los token en minúsculas.

- Quitar elementos no alfanuméricos:

    - Se emplea la función *isalnum* de Python para detectar caracteres no alfanuméricos y no incluirlos como tokens del documento.

- Obtener el lexema de las palabras (lemmatization):

    - Se emplea la función lemmatize de WordNetLemmatizer perteneciente a la librería NLKT. Mediante la obtención de lexemas, se representa de manera única conjuntos de palabras. De esta manera se aúna el significado semántico de palabras con mismo lexema.

Limpieza:

Para el análisis semántico, es intuitivo pensar que habrá conjuntos de palabras que no aportan valor semántico a los documentos. El proceso de limpieza se basa en eliminar las palabras menos relevantes, es decir, la que menos información aportan. Por ejemplo, artículos, preposiciones o conjunciones son palabras de escasa relevancia.

Se emplea la lista *stopwords* que proporciona NLKT. Esta lista contiene palabras del inglés consideradas de escasa relevancia semántica, se emplea para comparar los tokens del documento y poder eliminar los tokens que coincidan con los de la lista.

Vectorización:

Este es el proceso mediante el cual se representan los documento en el formato que requiere el modelo de tópicos LDA a su entrada.

Para el proceso de vectorización se emplea la librería gensim. La función *Dictionary* perteneciente a esta librería permite construir un diccionario Python que contiene todos los tokens que aparecen en el corpus y les asigna un identificador.

A partir de este diccionario y de la función *doc2bow* se puede transformar cada documento a la representación necesaria para su procesado mediante algoritmos que implemente LDA. De esta forma se construye el corpus e el formato necesario para llevar a cabo modelado de tópicos mediante algoritmos que implementen LDA.

3.4. Modelado de tópicos

El algoritmo empleado para implementar LDA es el que proporciona la librería gensim. Como ya se explicó en el capítulo anterior, los algoritmos que implementa LDA dependen de parámetros predeterminados. Uno de los que no se ha mencionado anteriormente es el número de iteraciones.

LDA es un algoritmo iterativo, un mayor número de iteraciones mejora la convergencia del modelo. Para la elección de este parámetro se atiende a una relación de compromiso entre el coste computacional, que aumenta con el número de iteraciones, y la convergencia del modelo. El valor que se elije son 500 iteraciones.

Para la selección del resto de parámetros se realizan las siguientes pruebas.

Hiperparámetros α y β de la distribución multivariada de Dirichlet:

Se fija en 5 el número de tópicos para analizar el efecto de α y β en el modelo generado para los dos corpus evaluados. Se comienza con los valores por defecto para ir aumentando progresivamente los valores de α y β. El criterio para la selección definitiva del valor de estos parámetros se explica en el siguiente capítulo.

Número óptimo de tópicos:

Con los valores de α y β ya establecidos se construye un modelo LDA para cada uno de los valores de $k \in [3,15]$. Una vez construidos los modelos se analiza la medida propuesta por Arun et al.[17] para determinar el número de tópicos óptimo.

Cabe mencionar que para la selección de valores k, se tiene en consideración que menos de tres tópicos es un modelo demasiado generalista para la muestra. A medida que el número de tópicos aumenta la interpretación temática de los mismos disminuye, por eso ante evaluaciones de colecciones de 10 blogs distintos se decide no evaluar más de 15 tópicos.

Se eligen los dos mejores de k para cada corpus, se compararán los resultados ofrecidos por cada valor de k en cada corpus

Visualizaciones:

Se plantean dos tipos de visualizaciones:

- Visualización de la composición de los tópicos mediante la herramienta *LDAvis.*

- Visualización de la distancia entre documentos vía MDS aplicando una como métrica de distancia la versión simétrica de la divergencia Kullback-Leibler.

    - Para esta visualización se emplea la librería de visualización Bokeh.

Para el análisis de resultados, solo la primera de estas visualizaciones será realmente significativa. Mediante la inspección visual de la composición, relación e importancia de los tópicos se realiza un análisis cualitativo del funcionamiento del modelo LDA.

Además, se ofrecerán resultados de cómo quedan asignados, a su tópico principal, los distintos post de cada blog. El objetivo será hacer una primera aproximación a la caracterización de la colección.

3.5. Grafos y análisis de red

Para la construcción de grafos y el posterior análisis de redes se suceden los siguientes procesos:

- Construir matriz de adyacencia pesada empleando la métrica de similitud de divergencia Jensen-Shanon.

- Transformar esta matriz en una matriz dispersa, con el objetivo eliminar los datos poco relevantes en cuanto a parecido de documentos. Esto se traduce en pasar de un grafo totalmente conectado , cada nodo conectado a todos los demás con el peso correspondiente, a un grafo con menor densidad de enlaces. Este proceso se lleva a cabo mediante umbralización. Se explica a continuación:

    - La divergencia de Jensens-Shanon devuelve valores entre 0 y 1. Se elige un umbral a partir del cual considerar que dos documentos están conectados, por ejemplo 0.5. Para cada par de documentos que presentan un parecido menor a 0.5 se establece que su parecido es 0. Para cada par de documentos que presentan una similitud mayor a 0.5 se realiza una correspondencia lineal para asignar un nuevo valor de similitud siguiendo:

      Sea $similitud(d_i, d_j) = x > umbral$

      $$y = \frac{x - umbral}{1 - umbral}$$

      Donde y es el nuevo valor de similitud para los documentos (d_i, d_j)

- Construcción de grafos mediante empleando la librería *netwokx*. A partir de los grafos construidos detectar comunidades empleando la librería *community*, que implementa el método de Louvain sobre grafos de *networkx*.

- Por último se emplea el software de visualización Gephi para realizar el cálculo de la intermediación, medida de centralidad de red empleada, y visualizar los grafos que establecen las relaciones entre documentos

Capítulo 4

4. Presentación y Análisis de Resultados

En el presente capítulo se presentan los resultados de aplicar el modelo propuesto sobre las dos colecciones de blogs que conforman los conjuntos de evaluación.

Cabe mencionar que tanto *PyLDAvis* como la herramienta Gephi son herramientas interactivas. Para dar respuesta a la pregunta de investigación es necesario emplear estas dos herramientas. Por este motivo, en este capítulo se presentan resultados y ejemplos de visualizaciones, pero además, al final del capítulo se intentará reflejar un ejemplo de cómo se podría caracterizar una colección de blogs empleando *PyLDAvis* y Gephi.

4.1. Presentación de los conjuntos de evaluación

A continuación se exponen las URLs de los 15 blogs que componen cada colección, se incluye el tamaño del vocabulario detectado tras el procesado de textos en cada colección.

COLECCIÓN 1	COLECCIÓN 2
http://mouldywriting.blogspot.com.es/?expref=next-blog http://www.stereopills.com/?expref=next-blog http://mtjrrantsravesonmusic.blogspot.com.es/?expref=next-blog http://pithytitlehere.blogspot.com.es/?expref=next-blog http://weeebeeemeee.blogspot.com.es/?expref=next-blog http://thescreever.blogspot.com.es/?expref=next-blog http://rhymeswithplague.blogspot.com.es/?expref=next-blog http://www.nikolaysblog.com/?expref=next-blog http://fortlauderdaleseoexperts.blogspot.com.es/?expref=next-blog http://whizzywords.blogspot.com.es/?expref=next-blog&view=classic http://internet-marketing-views.blogspot.com.es/?expref=next-blog http://12dolphins.blogspot.com.es/?expref=next-blog http://marriedbabymamamistress.blogspot.com.es/?expref=next-blog http://tantra-ssage.blogspot.com.es/?zx=d3378a7f1117ba86 http://www.punnuspage.com/?expref=next-blog	http://reedrambles.blogspot.com.es/?expref=next-blog http://creekside1.blogspot.com.es/?expref=next-blog http://school.albaseerah.com/?expref=next-blog http://cougarblueforever.blogspot.com.es/?expref=next-blog http://carladaviddesign.blogspot.com.es/?expref=next-blog http://michaelsmillerphotography.blogspot.com.es/?expref=next-blog http://rodocreative.blogspot.com.es/?expref=next-blog http://americandesiredefined.blogspot.com.es/?expref=next-blog http://agshorsley.blogspot.com.es/?expref=next-blog http://ixia-wonderwallflower.blogspot.com.es/?expref=next-blog http://wisiansways.blogspot.com.es/?expref=next-blog&view=classic http://orrstreetseries.blogspot.com.es/?expref=next-blog http://www.lankauniversity-news.com/?expref=next-blog http://escapistsguidetolife.blogspot.com.es/?expref=next-blog http://blueteddyblog.blogspot.com.es/?expref=next-blog
Número de tokens=10,727	Número de tokens=8,490

Tabla 2. Enlaces a los Blogs que conforman cada colección bajo estudio

4.2. Elección de parámetros α y β

En la primera fase de la generación del modelo LDA, se estudia el efecto de los parámetros α y β sobre los resultados del modelo. Tras este proceso se decide establecer los siguientes valores:

- α=0.6
- β=0.01

El proceso que condujo a la selección de estos valores se explica a continuación.

En primer lugar, con un número de 5 tópicos preestablecido, se estudió la salida del modelo LDA para los dos conjuntos de evaluación con los valores por defecto de α(α=0.1) y β(β=0.01). Se observó que a la salida del modelo la mayoría de los documento eran asignados a los distintos tópicos con probabilidades superiores a 0.9. Esto quiere decir, que cada documento expresaba información de pertenencia a un único tópico.

Esta representación no era válida para las colecciones bajo estudio. Dada la heterogeneidad en cuanto a su contenido, una categorización de los documentos tan rígida no expresaba relaciones entre blogs, ni entre documentos.

Como se explica en el segundo capítulo, el valor de α, influye en como se asigna la pertenencia de un documento a los distintos tópicos. A valores más grandes, se considera que un documento está representando por la combinación de un mayor número de tópicos. De esta manera, si una colección de documentos no está claramente categorizada se recomienda el uso de valores de α cercanos a 1.

Ante pruebas con valores de α del orden de 0.8 y 0.9 la salida del modelo para los dos conjuntos de evaluación presentaba documentos en los que casi ningún documento pertenecía a un tópico concreto con probabilidad mayor que 0.5. Esto implica, que los documentos quedaban representados como una combinación de tópicos con parecido valor de pertenencia para cada uno. Esta representación tampoco permite caracterizar las colecciones bajo evaluación.

Lo que se pretende es tener una representación en un punto medio de los dos extremos antes expuestos. Es decir, tener documentos representados como una combinación de tópicos, pero en los que uno de los tópicos tenga una importancia significativa[7]. En este fase de la investigación se comprobó que la representación objetivo se conseguía con los valores antes expuestos.

Sobre el parámetro β, se deja su valor por defecto. Cabe mencionar que se probó el efecto de este parámetro sobre la representación. Como se explica en el segundo capítulo, valores más grande de β, aumenta el número de palabras que definen un tópico. A priori se podría pensar que sería necesario aumentar el valor de este parámetro, pero aumentar este parámetro se traduce en que los tópicos estuvieran formados por conjuntos de palabras relevantes muy similares. Esto se observó en la composición de los tópicos mediante *PyLDAvis*. Por este motivo se selecciono su valor por defecto.

4.3. Elección del número de tópicos

A continuación se presentan resultados de la evaluación de la medida propuesta para la selección del número de tópicos. También se presentan imágenes de la distribución de tópicos vía *PyLDAvis,* y de la distribución de documentos vía MDS. Para cada colección, se ofrecen resultados para los dos mejores y el peor número de tópicos.

[7] Como se explica más adelante, con importancia significativa se hace referencia a que uno de los tópicos esté representado en un documento con probabilidad mayor o igual a 0.5.

Colección 1:

Distribución de la medida de evaluación (eje Y) para el conjunto de tópicos [3,15] de evaluación (eje X)

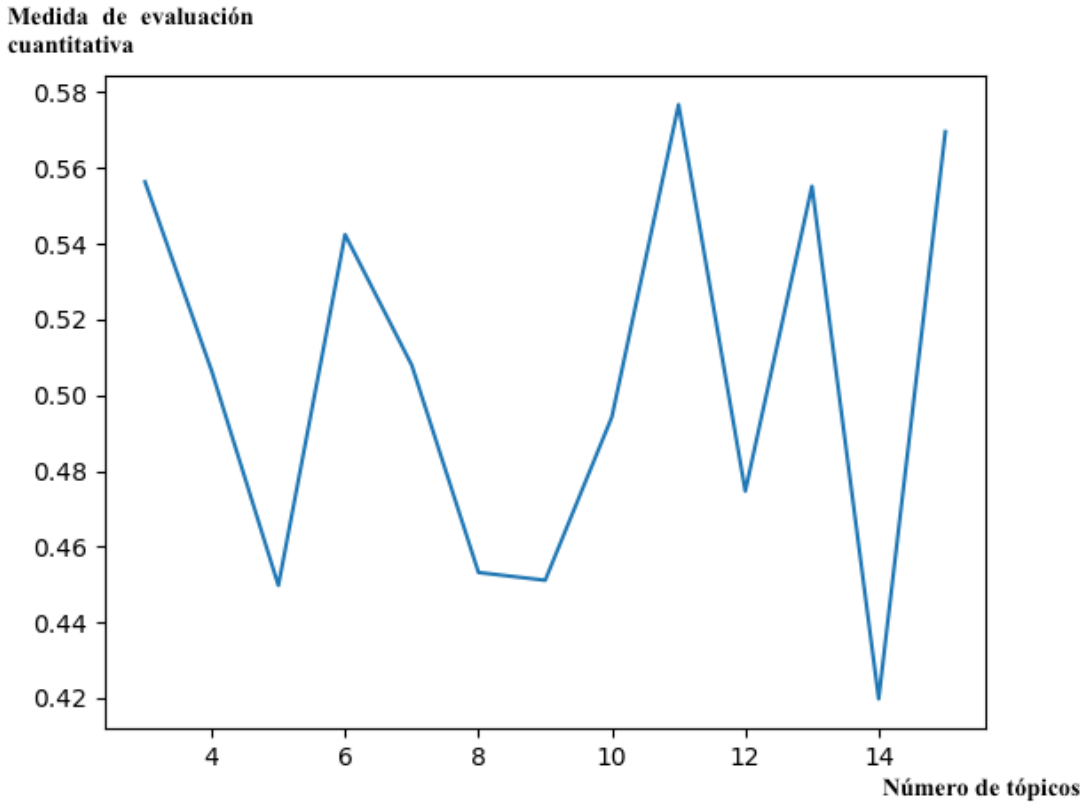


Figura 7: Distribución de la medida de evaluación (eje Y) para el conjunto de tópicos [3,15] de evaluación (eje X) para la colección 1

Como se puede observar los dos mejores valores son 14 y 5, siendo 11 el peor número de tópicos.

Cabe mencionar, que el color en la visualización de la distribución de documentos, indica el tópico con mayor asignación de probabilidad.

Para 5 tópicos:

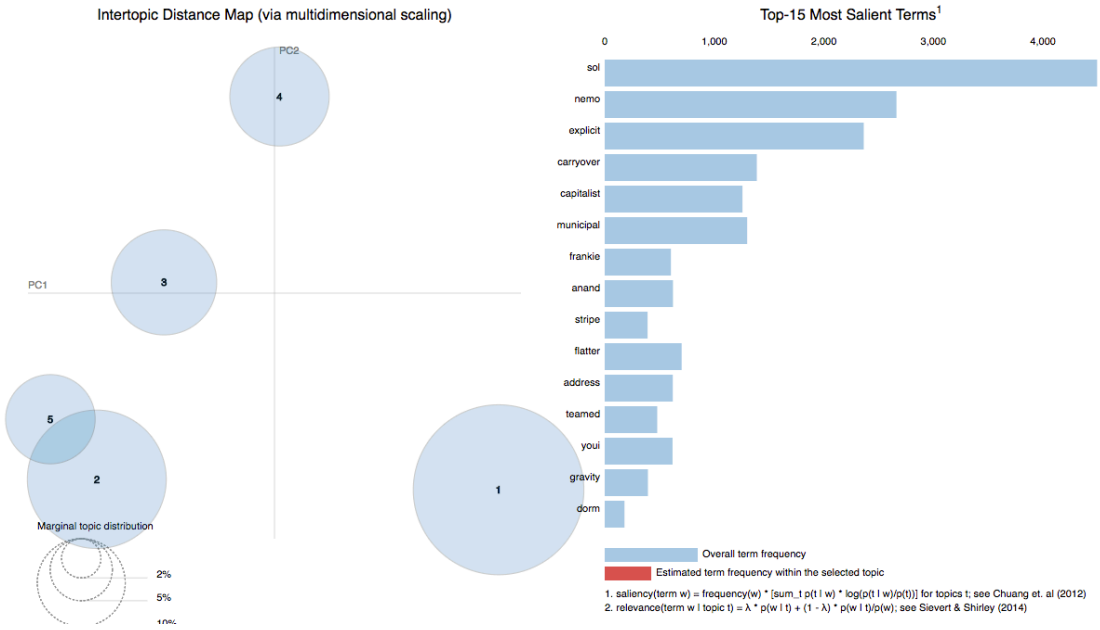


Figura 8: Visaulización de tópicos para modelo de 5 tópicos en la colección 1

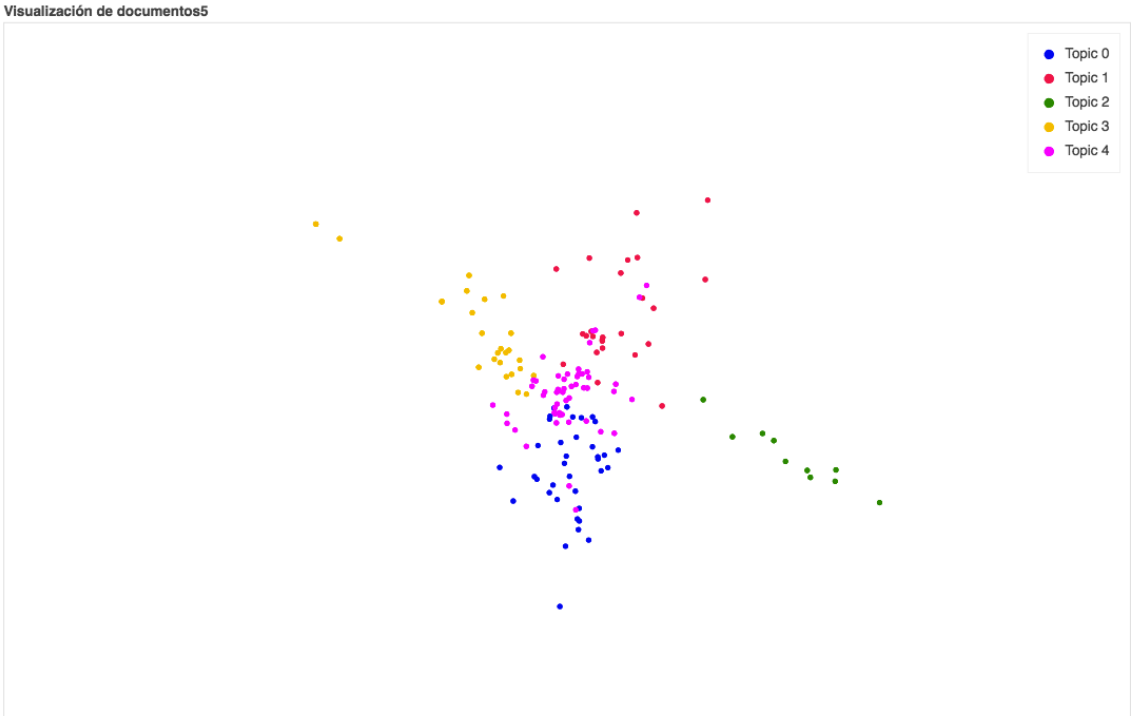


Figura 9: Visualización de documentos vía MDS para modelo de 5 tópicos en la colección 1

Para 14 tópicos:

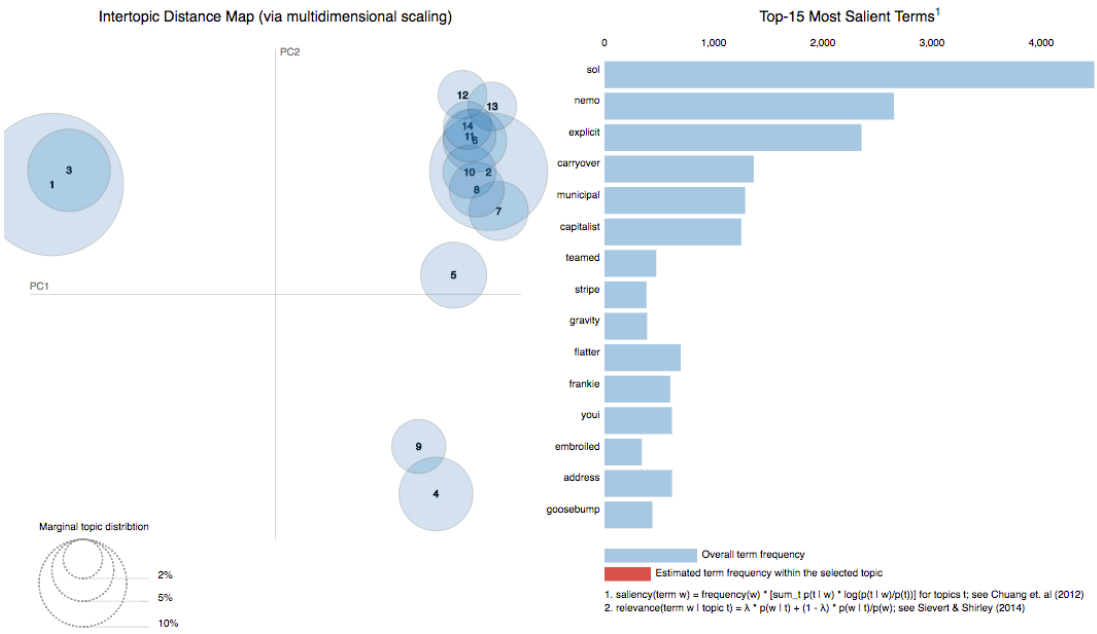


Figura 10: Visaulización de tópicos para modelo de 14 tópicos en la colección 1

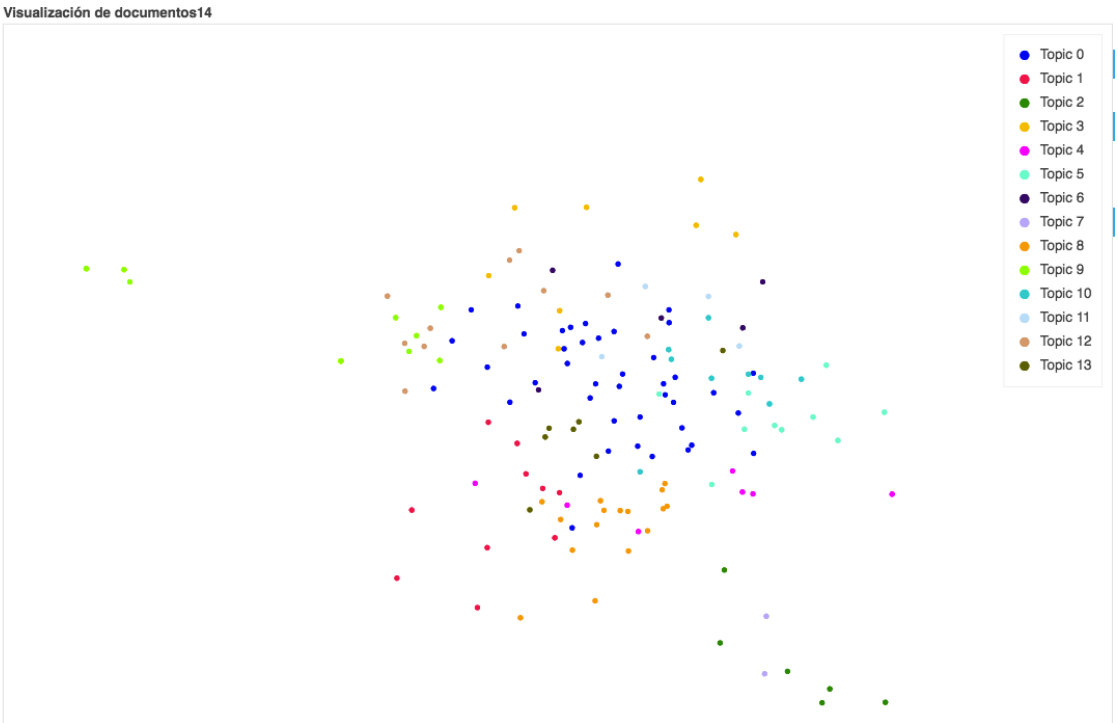


Figura 11: Visualización de documentos vía MDS para modelo de 14 tópicos en la colección 1

Para 11 tópicos:

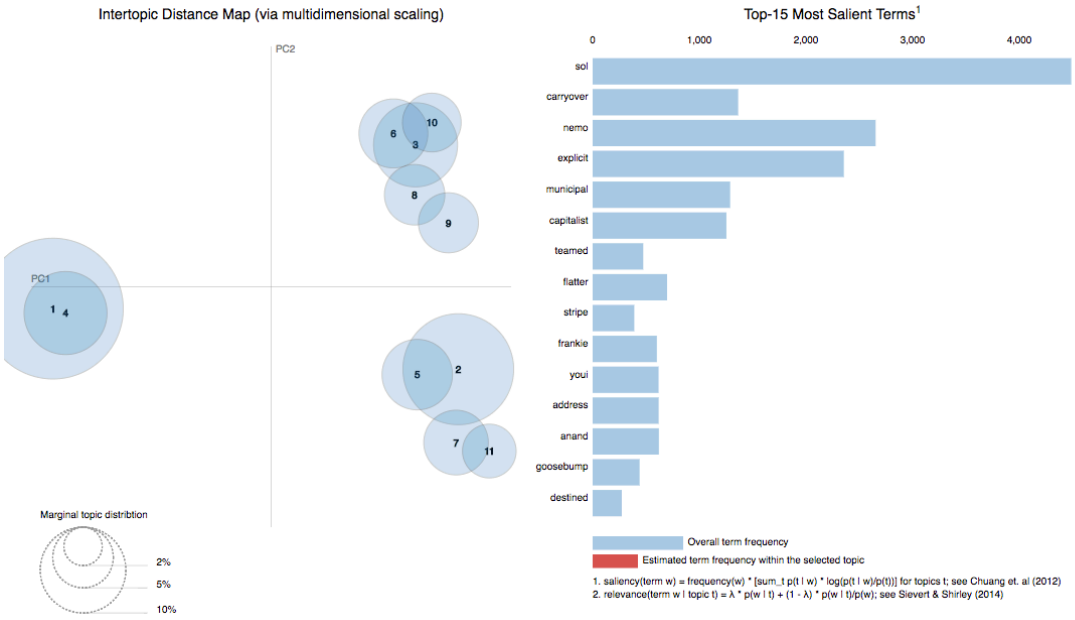


Figura 12: Visaulización de tópicos para modelo de 11 tópicos en la colección 1

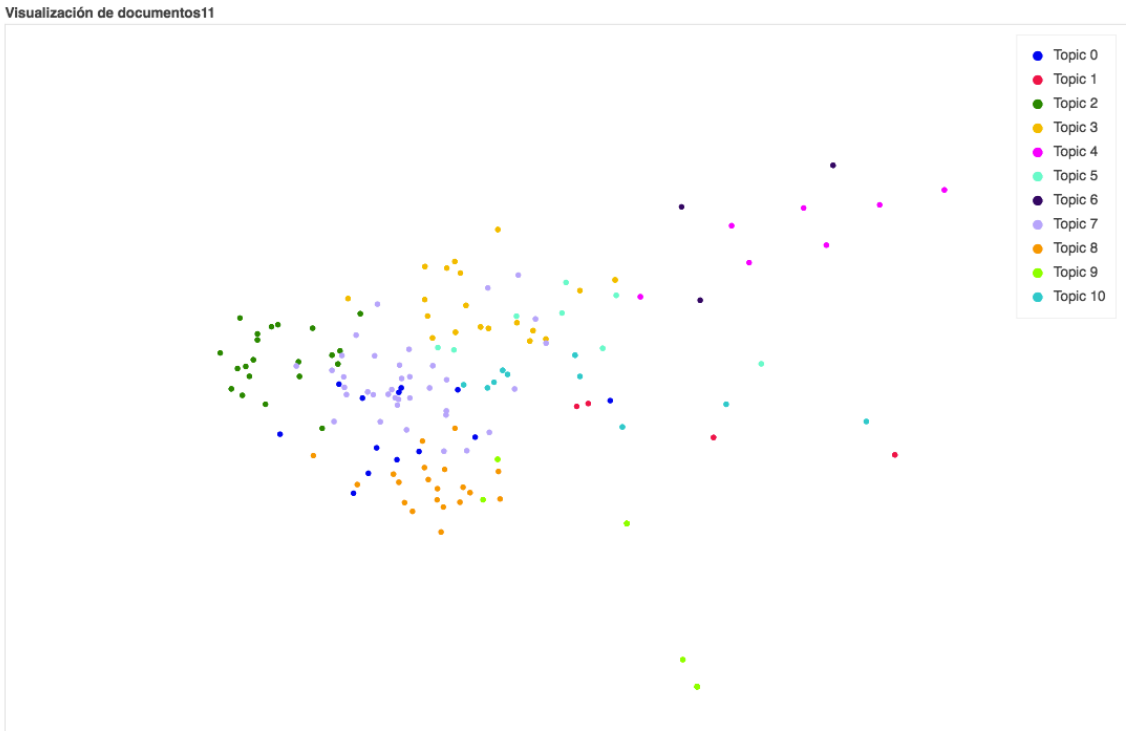


Figura 13: Visualización de documentos vía MDS para modelo de 11 tópicos en la colección 1

Analizando la composición de los tópicos mediante *PyLDAvis*, y de documento vía MDS se comprueba lo siguiente:

Para 5 tópicos:

Los tópicos tienen una importancia parecida dentro del corpus. Además, los tópicos no presentan grandes parecidos en su composición de palabras más relevantes. Destacar también, que en la distribución espacial de los tópicos[8], solo presentan parecido notable los tópicos 5 y 2.

En cuanto a la distribución de documentos, se aprecian 5 grupos bien diferenciados. Todo esto hace pensar que este modelo es más adecuado que el de 14 tópicos para caracterizar a esta colección pese al resultado de la medida de evaluación.

Para 14 y 11 tópicos:

En ambos casos, los tópicos tiene una importancia dispar dentro del corpus. Además, en ambos casos también hay dos tópicos principales en cuanto a relevancia en el corpus.

Se establecen agrupaciones de tópicos de poca relevancia que presentan una composición de palabras relevantes en las que se comparten algunos términos. Estos tópicos que representan a menor cantidad de documentos, están compuestos por palabras que presentan una Tf estimada dentro del tópico muy baja en comparación con los otros dos tópicos principales.

Todo esto induce a pensar que la colección no queda bien representada por un número de tópicos tan elevado.

En cuanto a la distribución de documentos, las agrupaciones de documentos pertenecientes a mismos tópicos se entremezclan. No se establecen grupos bien definidos y separados como en el caso de 5 tópicos.

Es evidente que a mayor número de tópicos más complejo se vuelve el modelo y la representación de documentos menos categorizada. A lo largo de este capítulo, y sobretodo en el siguiente, se establecen consideraciones que pretenden dar respuesta a estos comportamientos.

[8] Distancia inter-tópicos que calcula *PyLDAvis* y representa vía MDS

Colección 2:

Distribución de la medida de evaluación (eje Y) para el conjunto de tópicos [3,15] de evaluación (eje X)

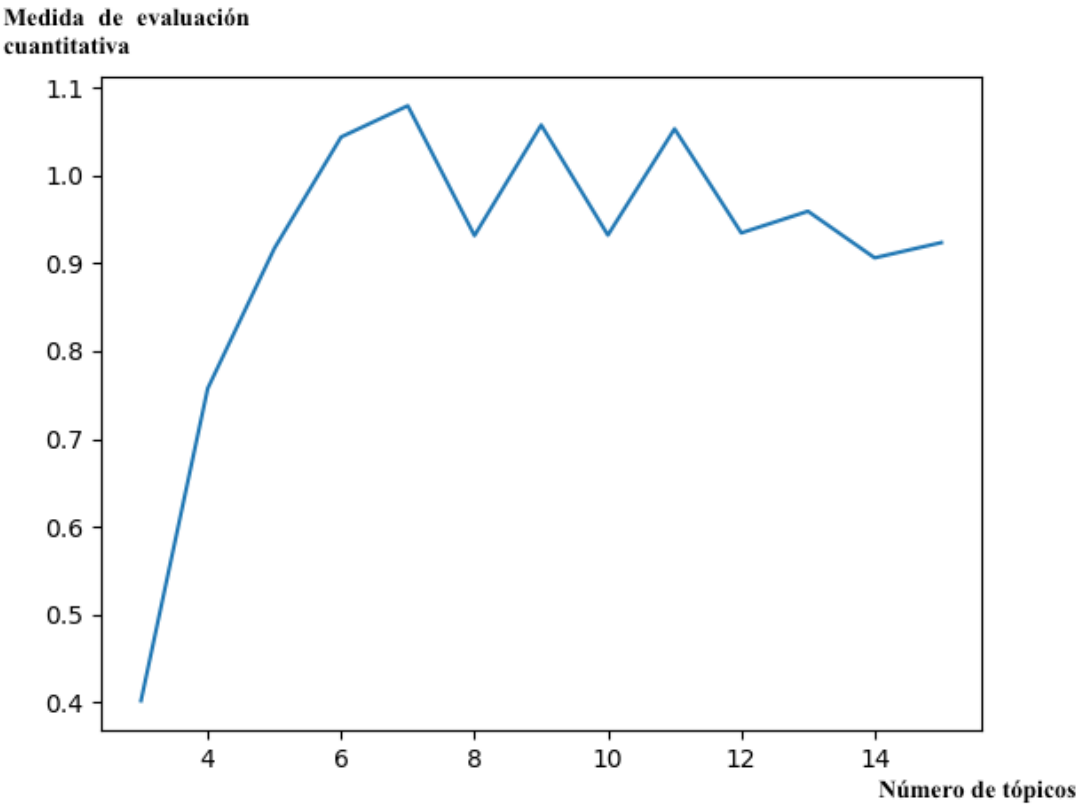


Figura 14: Distribución de la medida de evaluación (eje Y) para el conjunto de tópicos [3,15] de evaluación (eje X) para la colección 2

Como se puede observar los dos mejores valores son 3 y 4, siendo 7 el peor número de tópicos. En este caso se ve una clara tendencia creciente respecto a los dos primeros valores de prueba. La diferencia en el rango de valores entre los dos mejores tópicos y los demás es significativo en este caso. En la primera colección, el rango de variación es mucho menor.

Para 3 tópicos:

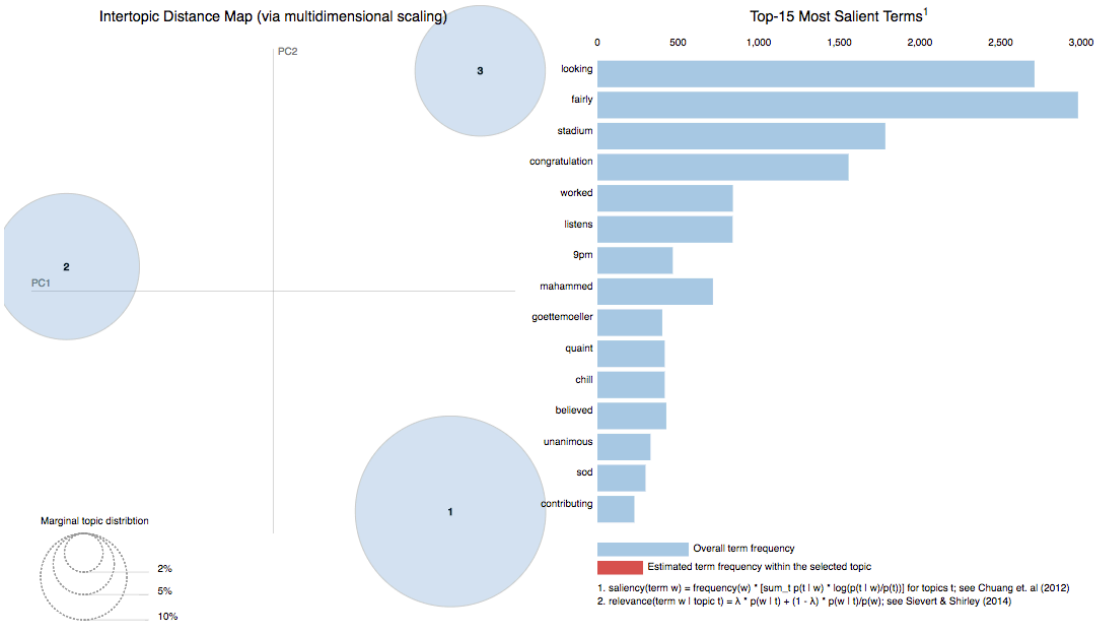


Figura 15: Visaulización de tópicos para modelo de 4 tópicos en la colección 2

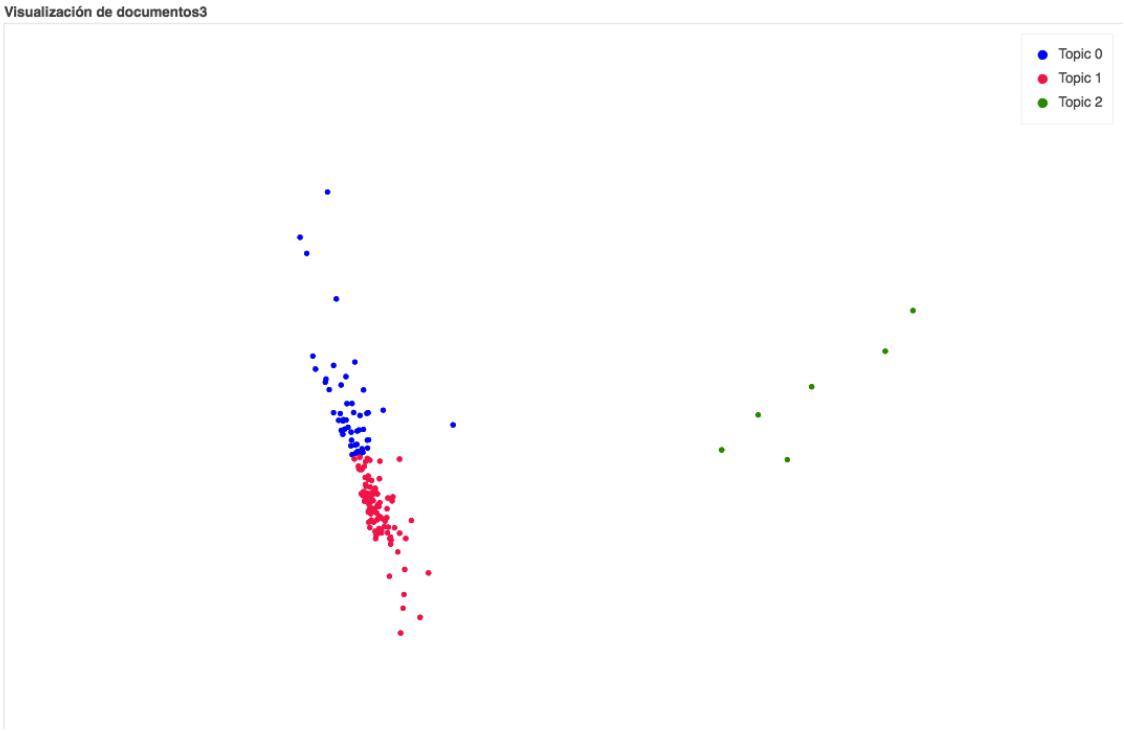


Figura 16: Visualización de documentos vía MDS para modelo de 3 tópicos en la colección 2

Para 4 tópicos:

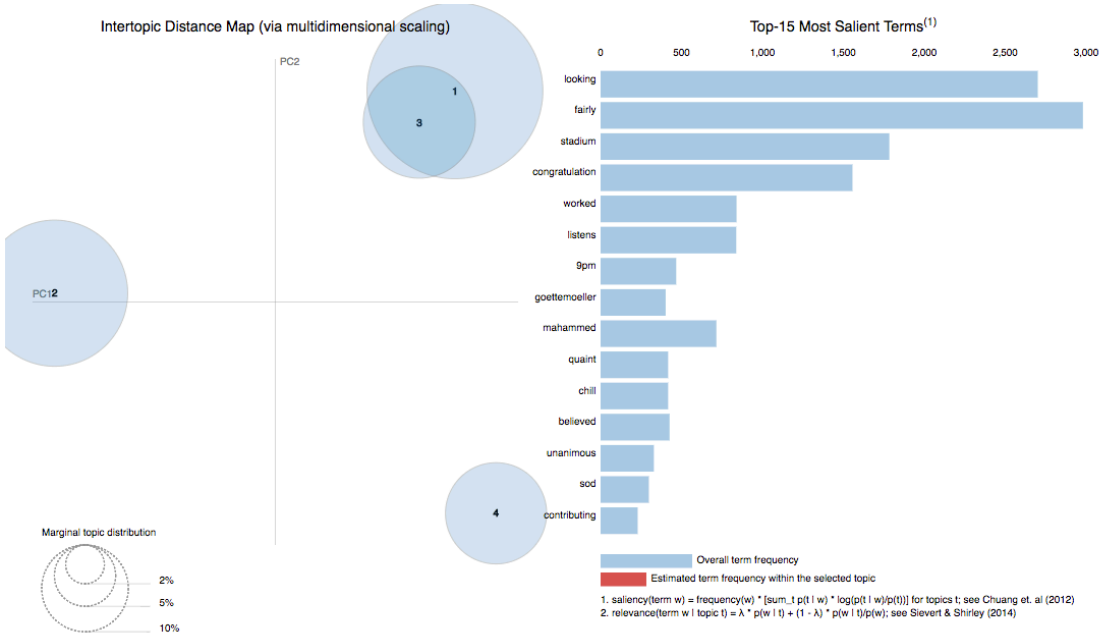


Figura 17: Visaulización de tópicos para modelo de 4 tópicos en la colección 2

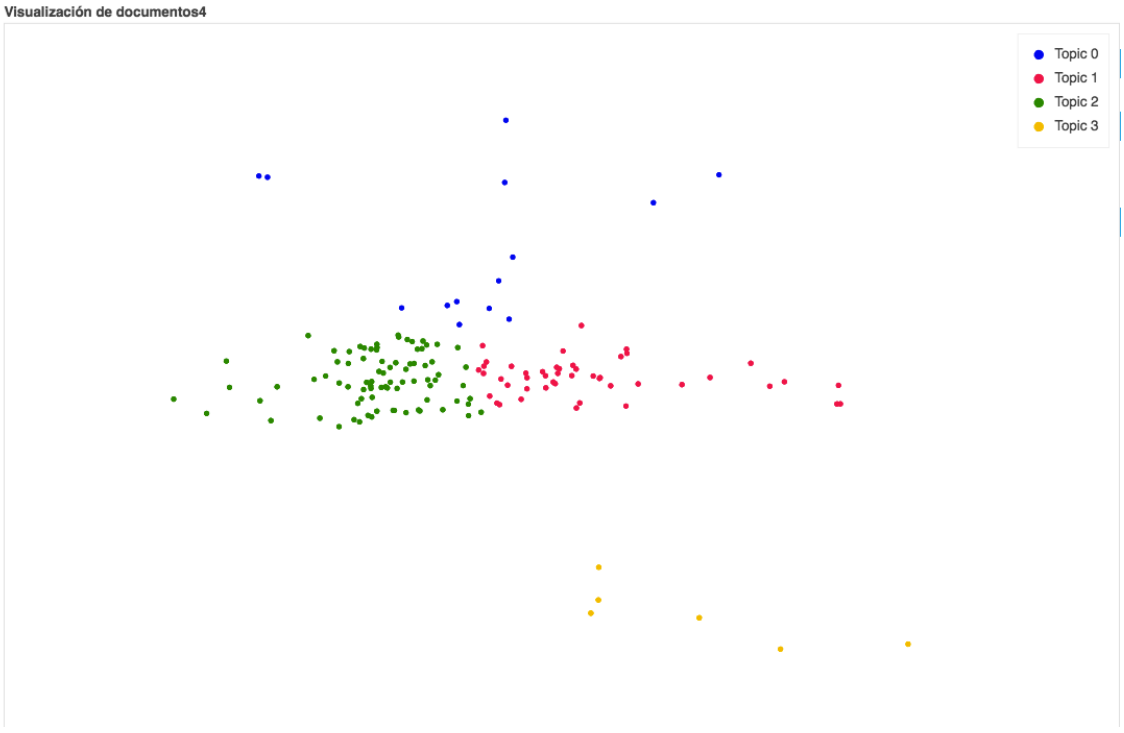


Figura 18: Visualización de documentos vía MDS para modelo de 4 tópicos en la colección 2

Para 7 tópicos:

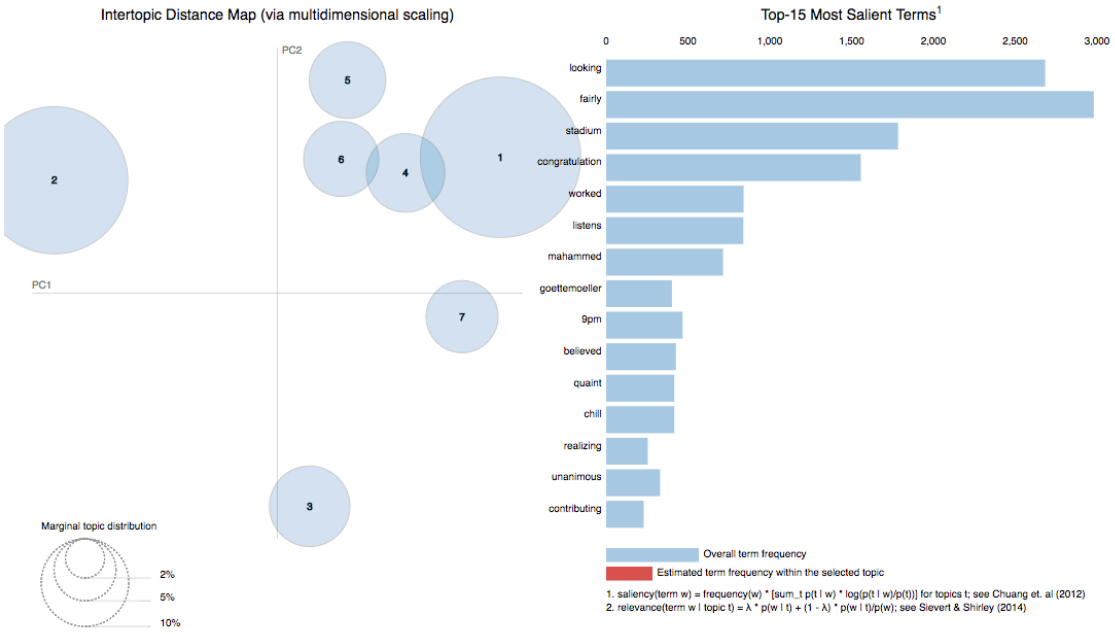


Figura 19: Visaulización de tópicos para modelo de 7 tópicos en la colección 2

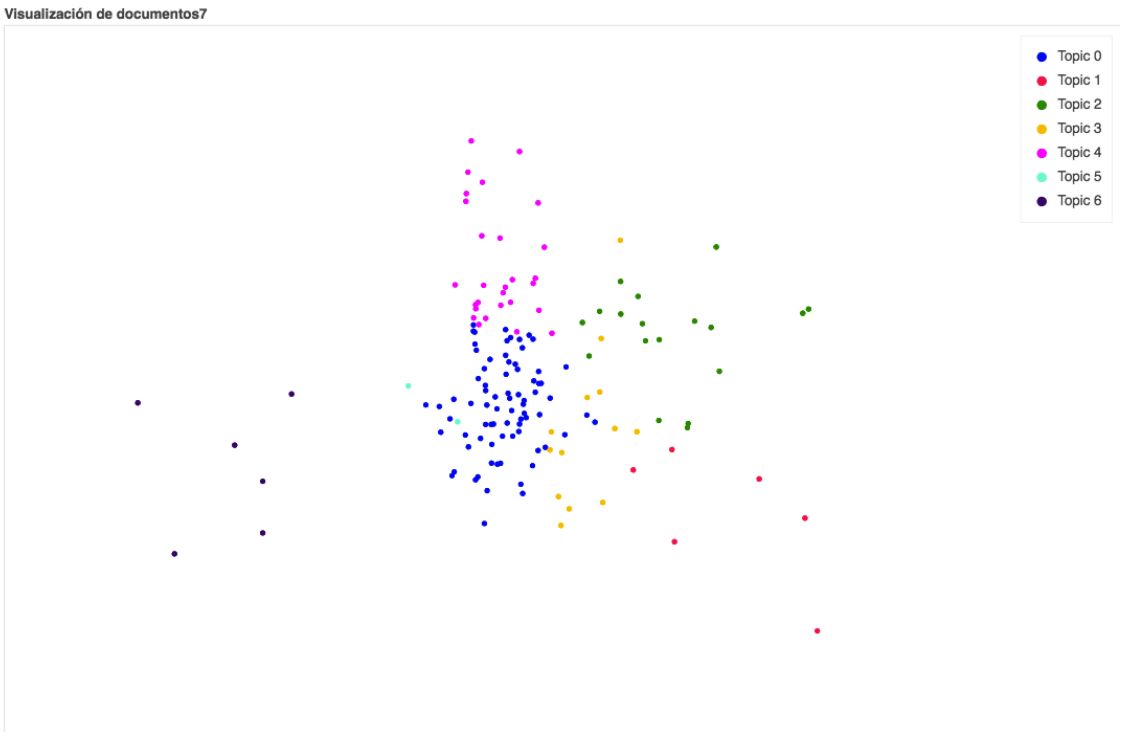


Figura 20: Visualización de documentos vía MDS para modelo de 7 tópicos en la colección 2

Analizando la composición de los tópicos mediante *PyLDAvis*, y de documento vía MDS se comprueba lo siguiente:

Para 3 tópicos:

Los tópicos presentan una importancia parecida dentro del corpus. Además, los tópicos no presentan grandes parecidos en su composición de palabras más relevantes. Destacar también, que en la distribución espacial de los tópicos[9], los tres tópicos se presentan alejados unos de otros a una distancia similar.

En la distribución de documentos se aprecian claramente tres grupos bien diferenciados.

Para 4 tópicos:

Se establece un tópico principal y otros tres de relevancia menor y similar. Además, los tópicos no presentan grandes parecidos en su composición de palabras más relevantes. Destacar también, que en la distribución espacial de los tópicos es muy similar a la del caso anterior, pero con el nuevo tópico establecido colindante al tópico principal.

En la distribución de documentos se aprecian claramente cuatro grupos bien diferenciados.

Esta situación induce a pensar que este modelo define mejor a un subgrupo perteneciente al tópico 3 del modelo con tres tópicos. Se ilustra y se explica a continuación:

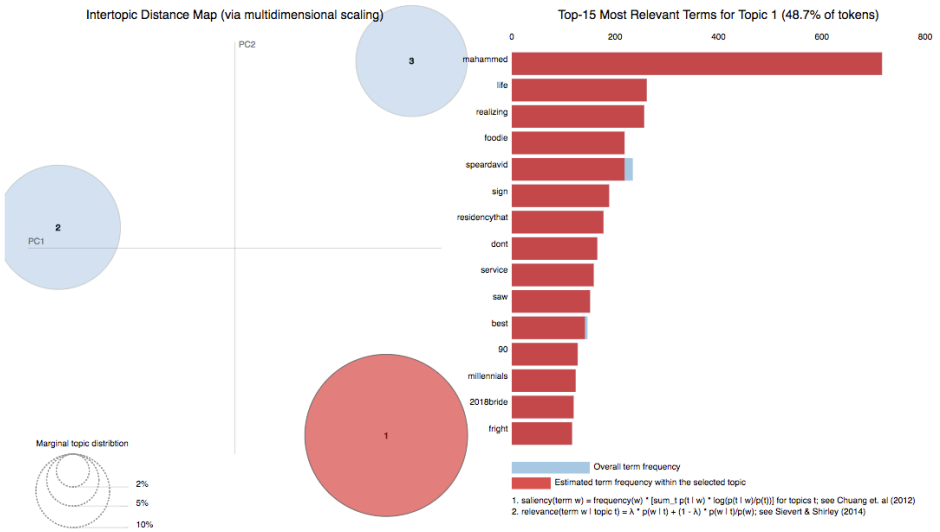


Figura 21: Visaulización de la composición del tópico 1 para modelo de 3 tópicos en la colección 2

[9] Distancia inter-tópicos que calcula *PyLDAvis* y representa vía MDS

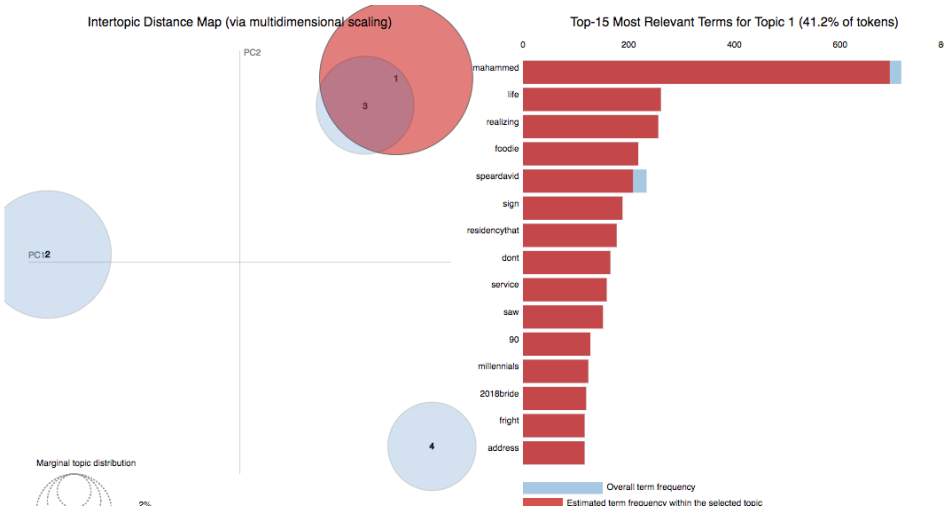


Figura 22: : Visaulización de la composición del tópico 1 para modelo de 4 tópicos en la colección 2

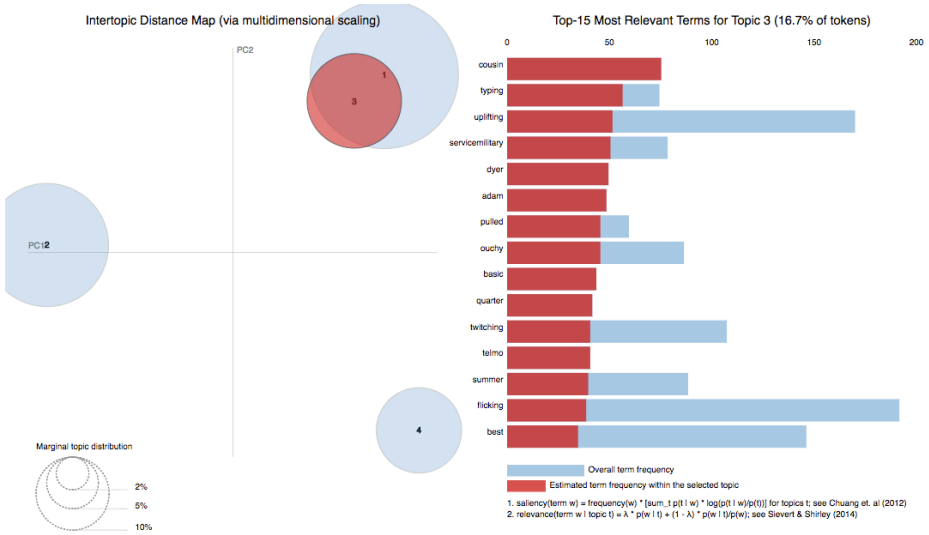


Figura 23: : Visaulización de la composición del tópico 3 para modelo de 4 tópicos en la colección 2

Como se puede observar, la composición del nuevo tópico es diferente, en sus palabras más relevantes, a la del tópico principal. Sin embargo, el tópico principal, es idéntico en relevancia y composición al tópico principal del modelo con 3 tópicos. A continuación se ilustra que el tópico 2 de los modelos es también idéntico.

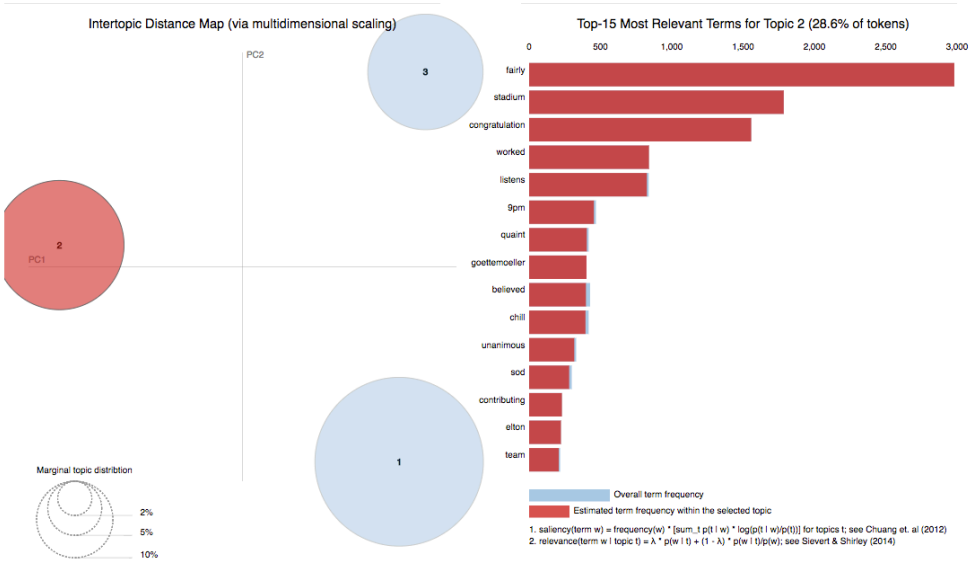


Figura 24: : Visaulización de la composición del tópico 2 para modelo de 3 tópicos en la colección 2

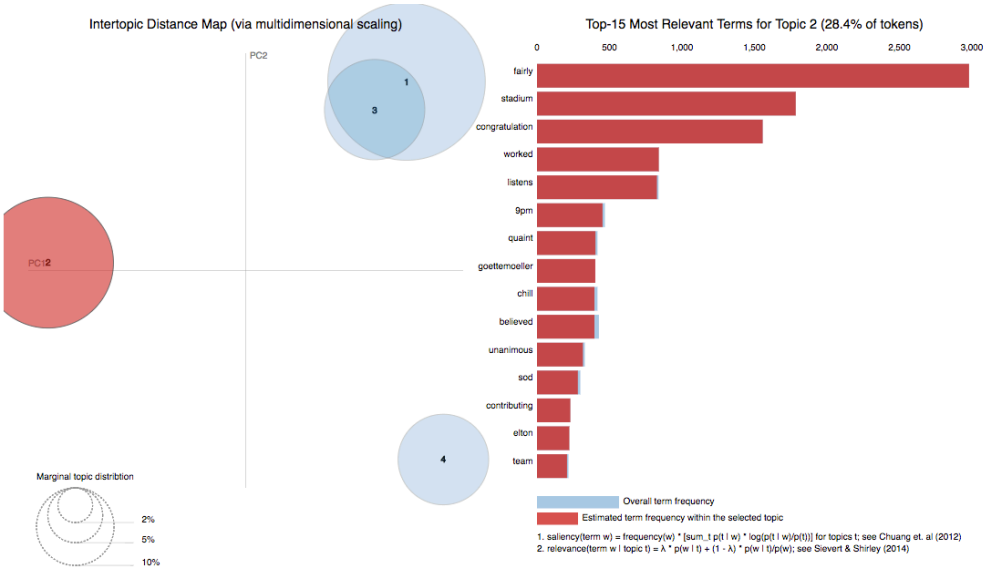


Figura 25: : Visaulización de la composición del tópico 2 para modelo de 4 tópicos en la colección 2

Existen evidencias cualitativas que indican que la colección 2 está mejor caracterizada mediante 4 tópicos que mediante 3.

Para 7 tópicos:

En este caso no se presentan problemas como en el la colección 1, esto es, tópicos compartiendo términos en los grupos de palabras más relevantes. La distribución de documentos permite distinguir a las distintas agrupaciones de tópico. El único inconveniente es que se establecen dos tópicos principales y el resto son de una relevancia baja y similar, representando a muy pocos documentos. Por tanto, es evidente que es una representación menos óptima para caracterizar la colección bajo estudio.

Como se explica en el siguiente apartado, tópicos de poca relevancia implica que no están representados de manera significativa en ningún documento.

4.4. Visualización de grafos y detección de comunidades

A continuación se presentan los resultados de la visualización mediante Gephi de los grafos construidos mediante las matrices de similitud. Se presentan grafos para los siguientes modelos:

Colección 1:

- Grafo para 5 tópicos
- Grafo para 14 tópicos

Colección 2:

- Grafo para 4 tópicos
- Grafo para 3 tópicos

Cabe mencionar que el valor de umbralización seleccionado para la construcción de las matrices de adyacencia es de 0.8. Este valor permite generar grafos no demasiado densamente conectados.

Las configuraciones empleadas para las visualizaciones mediante Gephi están disponibles en los anexos B.1, B.2, B.3 y B.4.

Colección 1:

Grafos para 5 tópicos

- Número de comunidades detectada=5

A continuación se ilustran visualizaciones del grafo para 5 tópicos. En la primera figura que se presenta, el color de los nodos indica el tópico más relevante. En la segunda figura, el color indica la comunidad de las detectada vía método de Louvain .

La etiqueta de cada nodo representa, mediante notación alfabética, la pertenencia a un blog. El número identifica la publicación dentro del blog. El tamaño de los nodos depende de la intermediación, calculada con Gephi, para cada nodo.

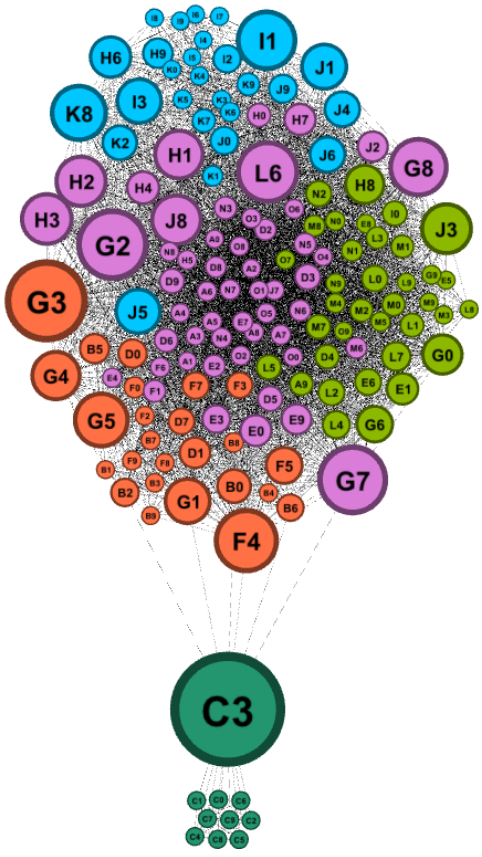


Figura 26: Visualización de grafos vía software Gephi para modelo de 5 tópicos en la colección 1. Color representa el tópico más relevamte

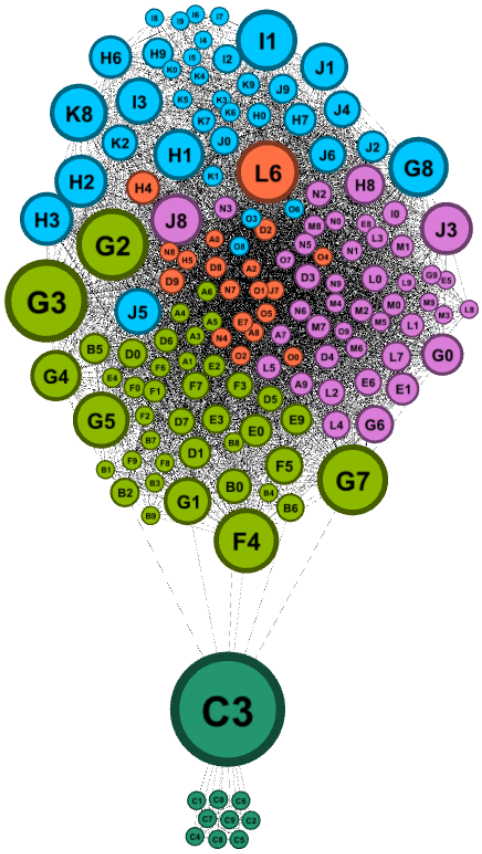


Figura 27 Visualización de grafos vía software Gephi para modelo de 5 tópicos en la colección 1. Color representa comunidad de pertenencia

Como se puede observar en las figuras 27 y 28, existe correspondencia entre las comunidades detectadas y las formadas a partir del tópico más relevante. Este resultado es un comportamiento esperado, dada la parametrización de α y β. Si recordamos lo que se explica al principio del capítulo, se seleccionan α y β para tener documentos representados como una combinación de tópicos, pero con uno de los tópicos con una importancia significativa. De esta manera, los documentos pertenecientes a un tópico de manera significativa se parecerían entre ellos.

En este punto cabe introducir los siguientes conceptos:

- Documento significativo: Documento en el que la pertenencia a uno de los tópicos queda representada por un valor mayor o igual a 0.5.

- Porcentaje de documentos significativos=$\frac{Documnetos\ significativos}{Tamaño\ de\ la\ colección}$

Para este caso encontramos:

$$\text{Porcentaje de documentos significativos} = \frac{109}{150} = 0.726$$

Mediante este modelo con 5 tópicos, obtenemos que en torno al 73% de los documentos son significativos. Lo más interesante es que los documentos no significativos, se corresponden con los que presentan los valores más altos de intermediación. Si un documento no es significativo, significa que no encaja en un tópico concreto, es más bien una combinación de distintos tópicos.

De esta forma se puede interpretar, en la visualización de grafos ,que nodos de mayor tamaño representan documentos de contenido menos específico. Pero a su vez, son los nodos centrales que conectan distintas agrupaciones.

Las visualizaciones mediante Gephi son interactivas, situándonos sobre un nodo concreto podemos visualizar a que nodos está directamente. Esto se ilustra a continuación. Se ilustra la subred que conforma un tópico que sus documentos significativos, son exclusivamente todas las publicaciones del Blog C.

En la figura 28 se aprecia claramente que dentro del Blog C, la publicación 3 es la única del Blog que presenta parecido con otras publicaciones de otros Blogs. Por este motivo su valor de intermediación es alto, puesto que es el nodo que conecta el al grafo principal con esta pequeña subred. Estudiar el contenido del documento C3, y el de los documentos a los que está conectado, puede ser útil para analizar como se relaciona el Blog C con los demás de la colección.

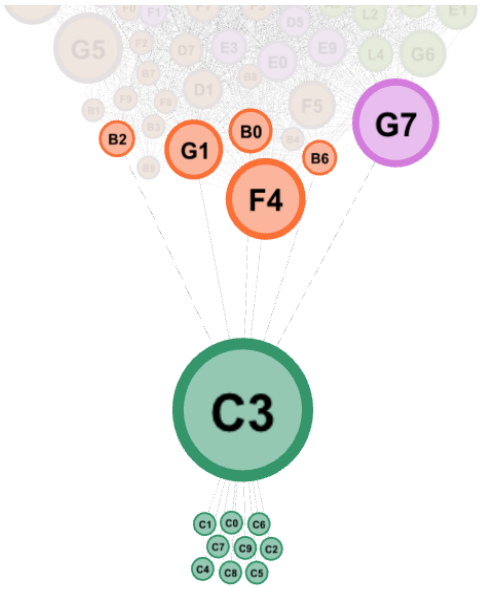


Figura 28: Visualización de grafos vía software Gephi para modelo de 5

Grafos para 14 tópicos

- Número de comunidades detectada=11

A continuación se ilustran visualizaciones del grafo para 14 tópicos. En la primera figura que se presenta, el color de los nodos indica el tópico más relevante. En la segunda figura, el color indica la comunidad de las detectada vía método de Louvain.

La etiqueta de cada nodo representa, mediante notación alfabética, la pertenencia a un blog. El número identifica la publicación dentro del blog. El tamaño de los nodos depende de la intermediación, calculada con Gephi, para cada nodo.

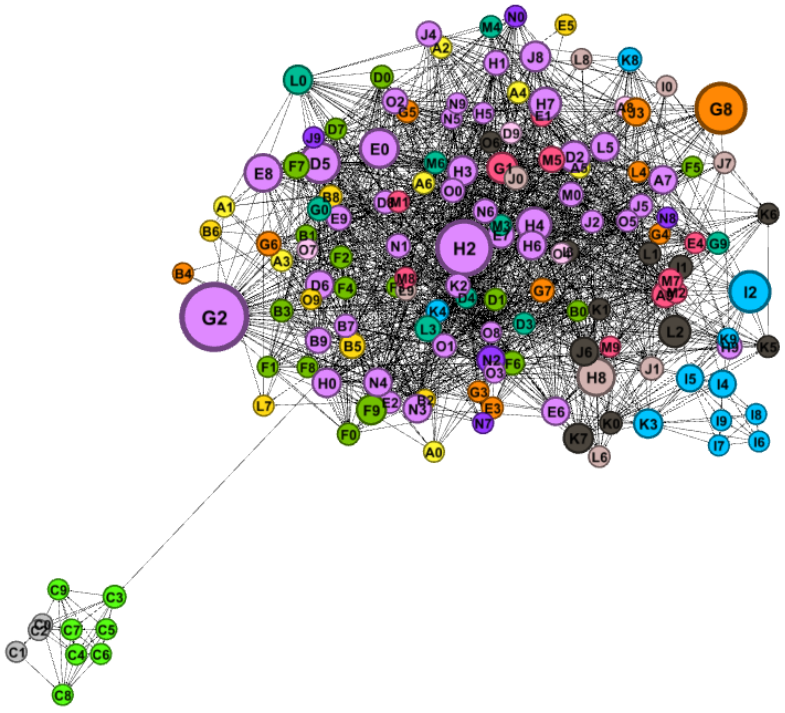


Figura 29: Visualización de grafos vía software Gephi para modelo de 14 tópicos en la colección 1. Color representa tópico más relevante

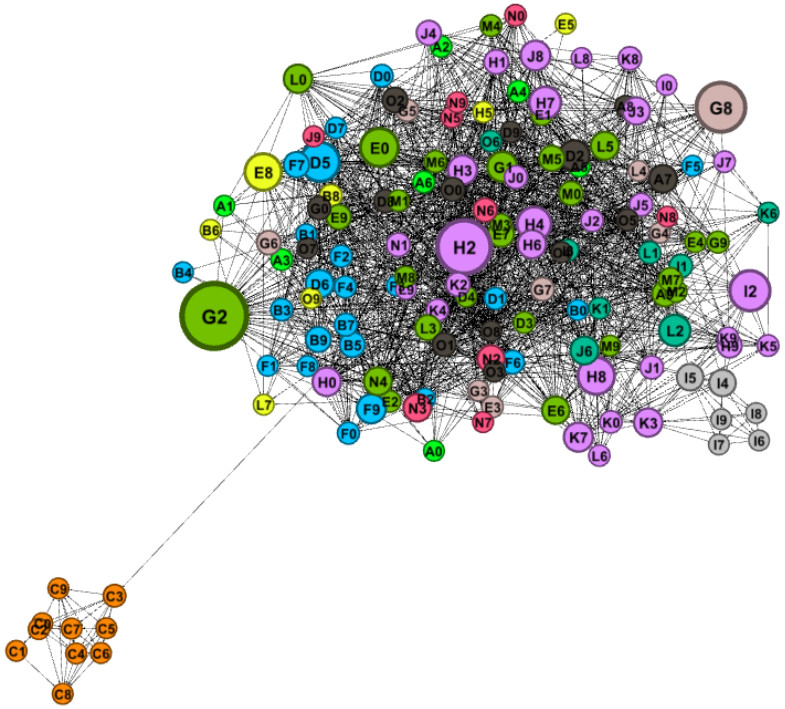


Figura 30: Visualización de grafos vía software Gephi para modelo de 14 tópicos en la colección 1. Color representa comunidad de pertenencia

Para este caso encontramos:

$$\text{Porcentaje de documentos significativos} = \frac{45}{150} = 0.3$$

Solo el 30% de los documentos son significativos, esto refuerza lo que se observó mediante *PyLDAvis,* este modelo de representación de la colección no es óptimo.

De la misma forma el número de comunidades detectadas es menor que el número de tópicos. Por ejemplo, la subred que conforman de nuevo las publicaciones del Blog C forman una comunidad. Sin embargo, no se agrupan todos en el mismo tópico principal como ocurría en el modelo de 5 tópicos.

Cabe mencionar que para este número de tópicos no se cumplen los criterios que se buscaban con la parametrización de α y β. En el siguiente capítulo se analiza con detalle esta situación.

Colección 2:

Grafos para 4 tópicos

- Número de comunidades detectada=4

A continuación se ilustran visualizaciones del grafo para 3 tópicos. En la primera figura que se presenta, el color de los nodos indica el tópico más relevante. En la segunda figura, el color indica la comunidad de las detectada vía método de Louvain .

La etiqueta de cada nodo representa, mediante notación alfabética, la pertenencia a un blog. El número identifica la publicación dentro del blog. El tamaño de los nodos depende de la intermediación, calculada con Gephi, para cada nodo.

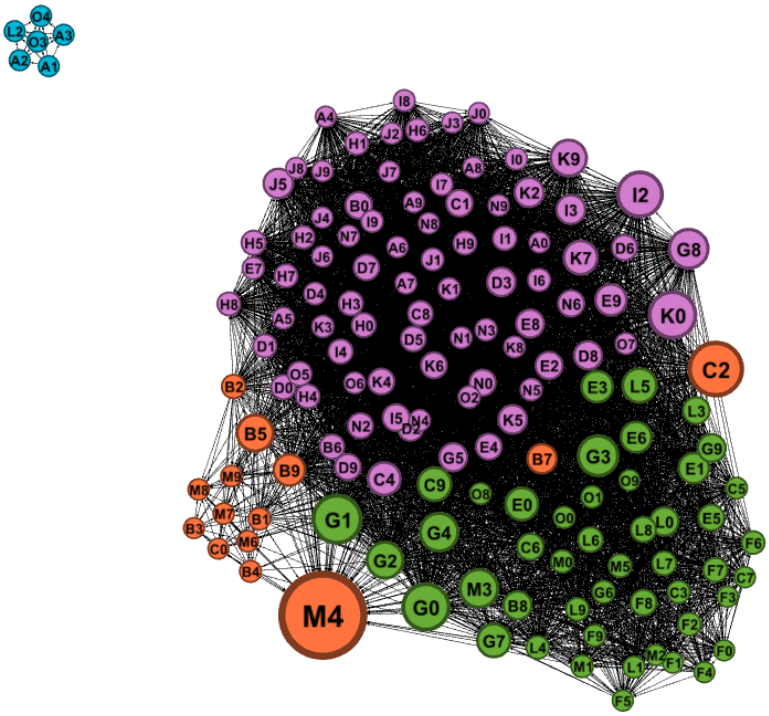


Figura 31: Visualización de grafos vía software Gephi para modelo de 4 tópicos en la colección 2. Color representa tópico más relevante

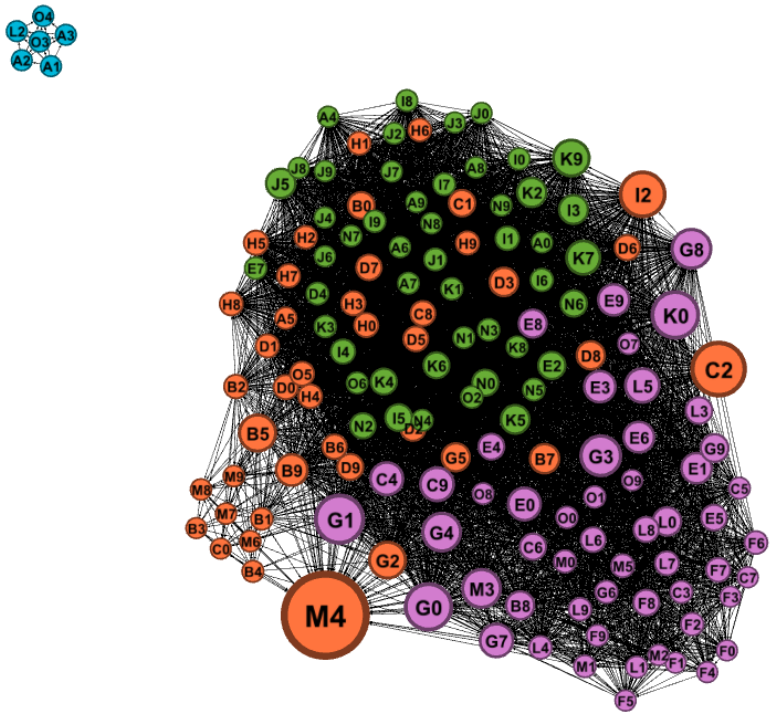


Figura 32: Visualización de grafos vía software Gephi para modelo de 4 tópicos en la colección 2. Color representa comunidad de pertenencia

Para este caso encontramos:

$$\text{Porcentaje de documentos significativos} = \frac{135}{150} = 0.9$$

Mediante este modelo, el 90% de los documento son representados como significativos. Se comprueba también que los documentos no significativos se corresponden con los de valores más altos de intermediación, nodos más grandes en las figuras 32 y 32.

Analizando las conexiones del nodo M4 mediante Gephi, como se ilustra a continuación, podemos observar que es un nodo que conecta la comunidades naranja y morada de la figura 32. Analizando el contenido de este nodo, y de nodos de pequeño tamaño directamente conectados, pertenecientes a ambas comunidades puede extraerse información que permita caracterizar a estas comunidades y por tanto, a la colección de blogs.

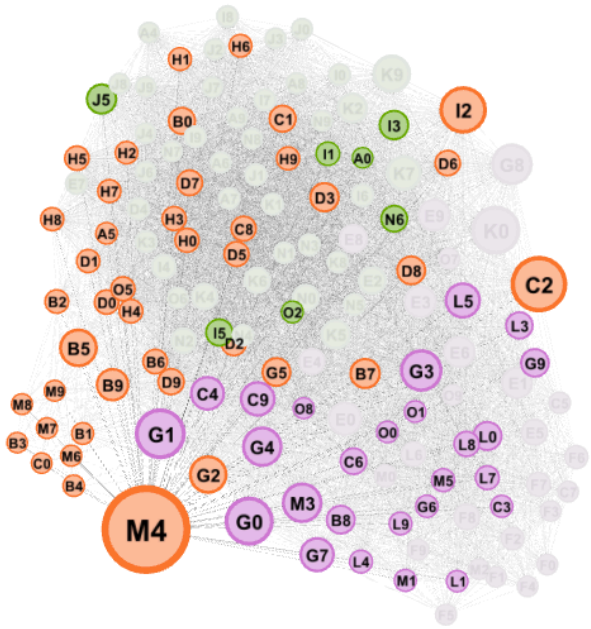


Figura 33: Visualización de grafos vía software Gephi para modelo de 4 tópicos en la colección 2. Color representa comunidad de pertenencia. Se resaltan conexiones del nodo M4

Cabe mencionar que existe elevada correspondencia entre agrupaciones de tópico más relevante y las comunidades detectada.

Destacar también, que en este grafo existe una subred en estrella que no está conectada al grafo principal. Esta subred representa una comunidad, y a un tópico concreto sin representación en el grafo principal. El estudio del contenido de estas publicaciones A1, A2, A3, L2, O3 y O4 puede revelar información útil para la caracterización de la colección.

Este tipo de comportamiento que se han ido detallando para este modelo, y para el modelo de 5 tópicos de la colección anterior, son los que son susceptible de aportar información significativa sobre la colección. Al final del capítulo se expone un ejemplo, de análisis cualitativo de estos comportamientos, apoyados en la lectura de documentos concretos, *PyLDAvis* y los grafos.

Grafos para 3 tópicos

- Número de comunidades detectada=3

A continuación se ilustran visualizaciones del grafo para 3 tópicos. En la primera figura que se presenta, el color de los nodos indica el tópico más relevante. En la segunda figura, el color indica la comunidad de las detectada vía método de Louvain.

La etiqueta de cada nodo representa, mediante notación alfabética, la pertenencia a un blog. El número identifica la publicación dentro del blog. El tamaño de los nodos depende de la intermediación, calculada con Gephi, para cada nodo.

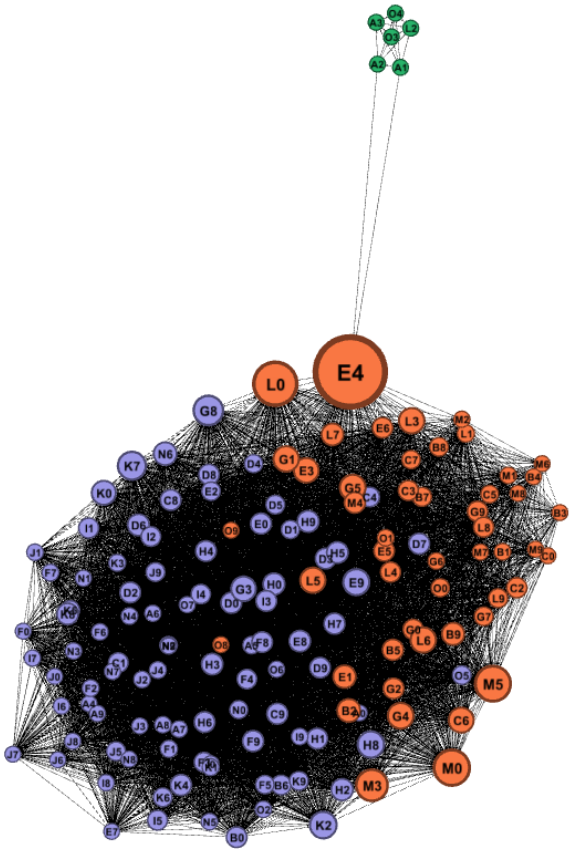


Figura 34: Visualización de grafos vía software Gephi para modelo de 3 tópicos en la colección 2. Color representa tópico más relevante

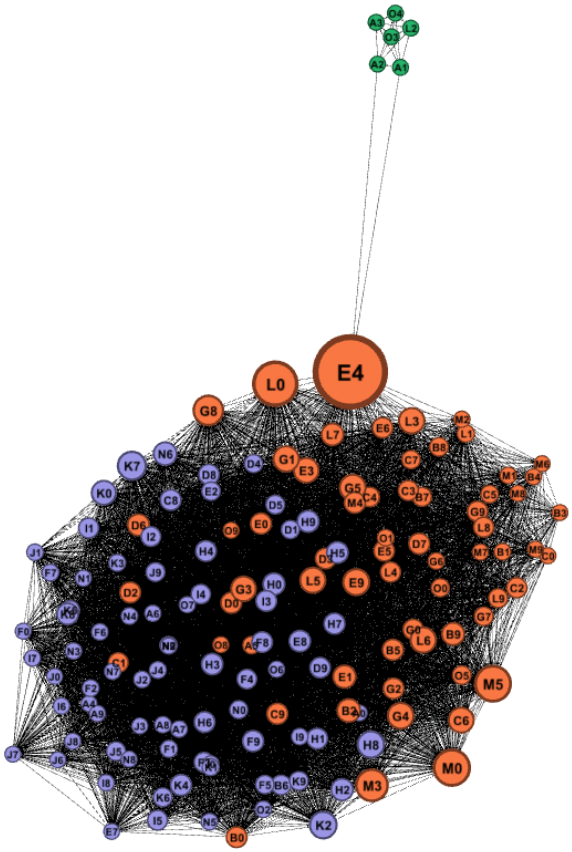


Figura 35: Visualización de grafos vía software Gephi para modelo de 3 tópicos en la colección 2. Color representa comunidad de pertenencia

Para este caso encontramos:

$$\text{Porcentaje de documentos significativos} = \frac{149}{150} = 0.993$$

El único documento que no es significativo es E4, que como se observa en las gráficas 34 y 35 es el documento de mayor intermediación de colección. Observando los grafos puede observarse que existe correspondencia entre las comunidades detectadas y las agrupaciones de tópicos más relevantes. Sin embargo, analizando el grafo resultado se pueden extraer conclusiones similares a las extraídas mediante *LDAvis*. A continuación se presenta una figura que representa los nodos a los que está conectado E4.

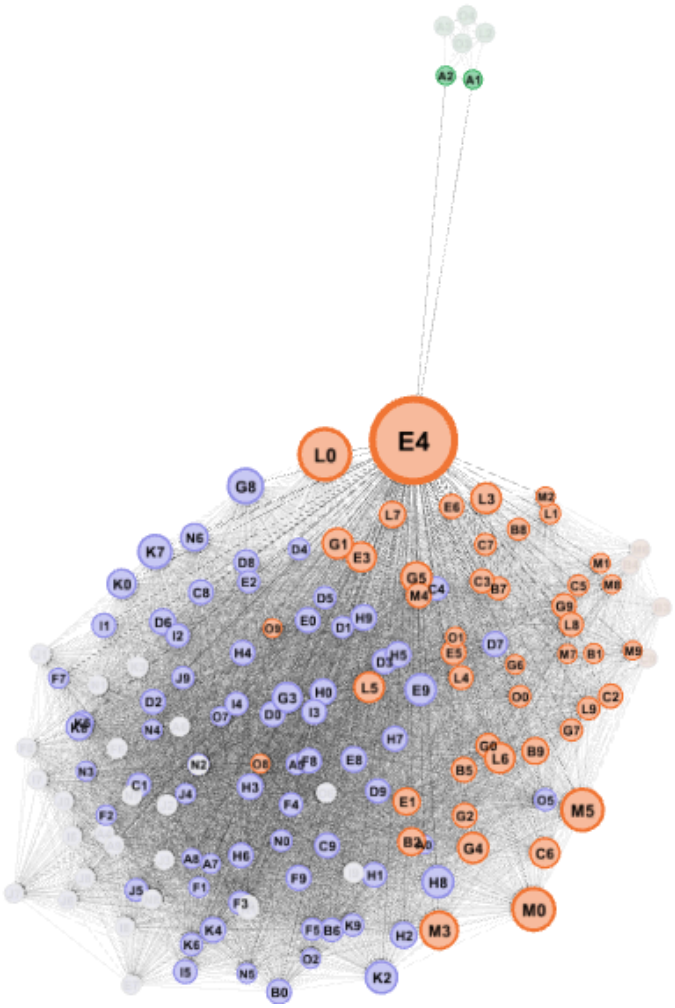


Figura 36: Visualización de grafos vía software Gephi para modelo de 3 tópicos en la colección 2. Color representa tópico más relevante. Se presentan las conexiones del nodo E4

Como se observa, el nodo E4 está relacionado con prácticamente todos los nodos de la red, por tanto no es lógico pensar que analizar su contenido pueda aportar información útil. Su valor alto de intermediación se debe a que conecta con el grafo principal la red que en el modelo de 4 tópicos estaba aislada. Esta situación no es comparable a la que se plantea en la colección 1 con el modelo de 5 tópicos y el nodo C3, figura 28. En ese caso C3 pertenecía a la subred, siendo el nodo que la conectaba al grafo principal, estando además conectado a un número mucho menor de nodos.

En el grafo del modelo de 4 tópicos, figuras 31 y 32, el nodo E4 presenta una intermediación muy baja, siendo uno de los 109 documentos significativos.

Esta situación conduce a pensar que el modelo de 4 tópicos es más óptimo para caracterizar esta colección.

Sin embargo, hay que resaltar que en cuanto a número de documentos significativos, solo los modelos de 5 y 4 tópicos, para las colecciones 1 y 2 respectivamente, cumplen con lo esperado atendiendo respecto a la parametrización de α y β. De hecho, los otros dos modelos representan las dos situaciones extremas que pretendían evitarse.

4.5. Ejemplo de uso combinado de PyLDAvis y visualización en grafo para caracterización de una colección de blogs

En primer lugar vamos a analizar la situación descrita para la primer colección con el modelo de 5 tópicos, en la que el nodo C3 conectaba al resto de nodos de ese blog con el grafo principal. Esto puede observarse en la figura 28.

Atendiendo al contenido de esta publicación, aportado en el anexo A.1,se ve claramente que trata sobre grupos musicales. Analizando el resto de publicaciones pertenecientes a la muestra de ese blog, se observa que todas tiene una estructura similar. El autor habla sobre nuevos discos y proyectos musicales, haciendo así una revisión de las novedades en el panorama musical.

En los anexos A.2 se puede encontrar el texto correspondiente a la publicación G1. De la misma manera el anexo A.3 muestra el texto correspondiente a la publicación F.4. Por último A.4, A.5 y A.6 contienen el texto de las publicaciones B0, B1 y B6 respectivamente.

Estas son las publicaciones a las que está conectado el nodo C3, todas ellas tienen temática musical. El documento G1 es una publicación el que el autor presenta la letra de una canción. Los documentos B0,B1,B6 y F4 son publicaciones con una temática de opinión y revisión de actualidad musical.

Analizando el contenido del resto de publicaciones de estos blogs se identifica rápidamente:

- El Blog F tiene un temática centrada en la opinión y revisión del género musical *punk.*

- El Blog B tiene una temática centrada en la opinión y revisión de música de origen neozelandés.

- El Blog G es un blog de opinión de actualidad general, en el que el autor en las publicaciones G1 y G7 trata temas relacionados con la música.

Analizando la composición semántica de los tópicos mediante PyLDAvis no se puede caracterizar el contenido de los distintos Blogs. La interpretación temática de los tópicos resulta realmente compleja, las agrupaciones de palabras no expresan una relación formal que permita intuir unidades temáticas interpretables.

La heterogeneidad de la muestra bajo estudio, en su temática y estilo, desemboca en que la caracterización de la composición de los tópicos aporte información difícilmente interpretable.

Sin embargo, la representación de publicaciones mediante estos tópicos expresados en un grafo, aporta información válida que permite caracterizar colecciones de blogs. Aunque, para esta caracterización, es necesario apoyarse en el contenido de las publicaciones para interpretar la información que aporta la visualización en grafo.

Sobre los análisis que se han realizado empleando conjuntamente la visualización de grafos y *PyLDAvis* cabe mencionar que se observa:

- La definición de tópicos es realmente sensible al estilo de redacción, ya que el lenguaje empleado va asociado al estilo.

- La definición de tópicos es sensible al tamaño de los documentos.

Por ejemplo en el caso que se está tratando en este capítulo, cabría preguntarse por qué si todos los documentos a los que está conectado C3 tratan de temática musical, no todos pertenecen al mismo tópico.

Las publicaciones del Blog C son extensas y tiene un estilo y lenguaje de escritura periodístico y técnico. Sin embargo, los otros dos blogs que tienen una temática sobre opinión musical están compuestos por publicaciones en lenguaje más coloquial. También las publicaciones de estos blogs son mucho menos extensas en comparación con las del Blog C.

Capítulo 5

5. Conclusiones

En el presente capítulo se presentan las conclusiones extraídas del análisis de resultados. Este capítulo tendrá dos apartados. Un primer apartado relativo a conclusiones sobre el modelo implementado, fundamentado en los resultados obtenidos. En un segundo apartado se aportan conclusiones de ámbito más general que atienden a responder la pregunta que motiva la investigación.

5.1. Conclusiones sobre el modelo implementado

Errores en el modelo implementado:

Mediante el análisis de los resultados proporcionados por el modelo, se ha detectado un error en el método para evaluar el número de tópicos que mejor define a una colección de blogs.

Al establecer los valores de α y β, con un objetivo de representación de documentos, se establecieron valores que cumplían para un número de 5 tópicos. Al optimizar el modelo para este número concreto, se desatendió como afectaban estos valores para el resto de número de tópicos. El efecto de estos de valores de α y β sobre números de tópicos mayores, es la pérdida progresiva de documentos significativos. Por el contrario, la disminución del número de tópicos se traduce en un aumento del número de tópicos.

Por este motivo no se pueden establecer evidencias de que los valores encontrados como mejores, 5 y 4 tópicos para las colecciones 1 y 2 respectivamente, sean los que mejor caracterizan a las colecciones bajo estudio. Sin embargo, ha podido comprobarse que estas representaciones caracterizaban adecuadamente ambas colecciones.

Por tanto, evaluar la medida propuesta por Arun et al.[17] para la selección del mejor número de tópicos, solo tenía sentido entre los modelos que cumplían con las condiciones de representación de documentos a l

Representación de documentos a la salida de un modelo LDA:

Gracias al error cometido en la planificación de los valores α y β, se ha podido comprobar la hipótesis de partida para a selección de los mismos. Si recordamos, lo que se busca es tener documentos representados como una combinación de tópicos, pero en los que uno de los tópicos tenga una importancia significativa.

Cuando la representación de documentos a la salida de un modelo LDA, no se asemeja, la extracción de información interpretable para la caracterización de colecciones de blogs es realmente compleja.

Visualización en grafo:

Sobre la visualización en grafo propuesta, se establecen las siguientes conclusiones:

- La intermediación aplicada como rango para el tamaño de los nodos, es útil para detectar nodos que conectan comunidades.

- Las relaciones que se visualizan en los grafos, combinado con el análisis del contenido de los documentos, expresan información útil para caracterizar los distintos blogs y también los propios tópicos descubiertos.

- La detección de comunidades es útil para generar agrupaciones de documentos similares y estudiar las relaciones que se establecen entre las mismas. Además, que el número de comunidades detectadas coincida con el de tópicos, y que existe correspondencia entre las agrupaciones de documentos por tópicos principales y las comunidades detectadas, ofrece información sobre el funcionamiento del modelo.

5.2. Conclusiones de ámbito general

¿Es el modelo LDA susceptible de aportar información, entendible a nivel humano que ayude a caracterizar tanto el contenido de los Blogs como las relaciones que se establecen entre los mismos?

Contestando a la pregunta de investigación, los datos a la salida de un modelo LDA pueden aportar información entendible y útil para caracterizar colecciones de blogs. Sin embargo, como se menciona al final del capítulo 4, la interpretación de la composición de los tópicos es realmente compleja y apenas aporta información sobre la temática de los blogs.

Por tanto, como se ha visto en el presente trabajo, para caracterizar colecciones de blogs a través de visualización de grafos es necesario recurrir a leer el contenido de ciertas publicaciones relevantes.

Ante corpus no categorizados y heterogéneos en cuanto a temática y estilo , el modelo LDA no ofrece información interpretable semánticamente de la composición de los tópicos. Por lo tanto, el modelo LDA no es suficiente para caracterizar una colección de blogs.

En el caso de que los tópicos expresasen unidades temáticas, combinándose con la visualización de grafos se podría caracterizar colecciones de blogs sin necesidad de recurrir a leer el contenido de las publicaciones.

Por último, como consecuencia del análisis de resultados y conclusiones sobre el modelo implementado, se propone un método para determinar el número de tópicos que mejor define un corpus desestructurado y heterogéneo:

1. Buscar, para cada número de tópicos bajo evaluación, un parametrización de α y β que permitan tener a la salida un porcentaje elevado de documentos significativos, pero no cercano al 100%. Justo buscando estos resultados, se encontrarán límites inferiores y superiores de números de tópicos que no cumplen para ningún valor de α y β.

2. Sobre el conjunto de modelos construidos, evaluar la medida propuesta por Arun et al.[17]

3. Emplear visualizaciones de tópicos, de grafos y detección de comunidades para analizar los modelos que ofrecen mejores resultados para la medida del paso anterior.

Capítulo 6

6. Líneas Futuras

En este capítulo y a modo de cierre del trabajo, se establecen las líneas futuras que pueden seguir a esta investigación, además se incluye un análisis del impacto del trabajo.

Sin duda una de las líneas futuras a plantearse es la mejora de la representación de datos a la entrada del modelo LDA. Es evidente, que ante corpus con gran variedad de temáticas y estilos, la representación de documentos mediante Tf es insuficiente.

Incluir mejoras como la técnica *Self-Term Expanssion*[24], de la que se habla en el capítulo 2, u otras que puedan enriquecer la representación de documentos.

Mejoras en estos términos ayudarían sin duda a que los tópicos descubiertos tuviesen una interpretación temática mucho más sencilla.

Otra de las líneas más interesantes está directamente relacionada con el estudio de redes de documentos. Como se ha visto en el desarrollo de este trabajo, la visualización de grafos permite obtener información de manera rápida que ayude a caracterizar colecciones de blogs.

Seguir avanzando en la aplicación de teoría de grafos sobre redes de documentos, puede ofrecer resultados interesantes. Por ejemplo, evaluando otras medidas de centralidad de red y su influencia en la interpretación de las visualizaciones. Todo con el fin de tener herramientas interpretables a alto nivel que ayuden a extraer información de corpus no categorizados.

En cuanto al impacto del trabajo, como se ha explicado en el primer capítulo, poder gestionar información expresada en lenguaje natural es uno de los grandes retos en el sector TIC.

Contar con herramientas que permitan caracterizar colecciones de blogs puede ser útil en áreas como el marketing o la sociología, pero además presenta un enorme potencial para ser útil en los sistemas de recomendación basados en contenido. En plataformas como Blogger, se podrían establecer sistemas de recomendación basados en contenido que implementarán visualizaciones de grafos y árboles.

Estas visualizaciones expresarían relaciones entre blogs parecidos en su estilo y temática. De esta forma, un usuario que reconociera un blog de su agrado podría encontrar blogs parecidos mediante una herramienta visual e intuitiva.

7. Bibliografía:

[1] J.M Roca, "¿Qué es la Sociedad de la Información?", *Informe TIC Fácil.* [En línea]. Disponible en: http://www.informeticplus.com/que-es-la-sociedad-de-la-informacion

[2] J.M Roca, "¿Qué son las TIC?", *Informe TIC Fácil.* [En línea]. Disponible en: http://www.informeticplus.com/que-son-las-tic

[3]Fuente en línea. Disponible en: https://es.wikipedia.org/wiki/World_Wide_Web

[4] Fuente en línea. Disponible en: https://definicion.de/informacion/.

[5] Fuente en línea. Disponible en: http://www.vicomtech.org/t4/e11/procesamiento-del-lenguaje-natural

[6] Hernández, D. Tomás y B.Navarro, "Una aproximación a la recomendación de artículos científicos según su grado de especificidad", *Procesamiento del Lenguaje Natura*l, Revista nº 55, páginas 91-98, 2015.

[7] D.M. Blei, A.Y. Ng, y Jordan.M. "Latent dirichlet allocation", *the Journal of machine Learning research*, páginas 993–1022, 2003.

[8]G.M. Salton, Wong, A y C.S. Yang "A Vector Space Model for Automatic Indexing". *Communicatons of the ACM*, páginas613-620, 1975.

[9] L. Arco, "Agrupamiento basado en la intermediación diferencial y su valoración utilizando la teoría de los conjuntos aproximados", Universidad Central Marta Abreu de las Villas, Cuba, 2008

[10] R. Moya. "Mapas gráficos para la visualización de relaciones en sistemas de recomendación", Trabajo fin de máster, Dpto de Sistemas Infrmáticso, Univesidad Politécnica de Madrid, Madrid, España 2015, 52-63.

[11] A. Chaney y D.M. Blei. "Visualazing Topic Models", presentado en Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, 2012.

[12] F. Guerrero y J.M. Ramírez. “El análisis de Escalamiento Multidimensional: Una alternativa y un complemento a otras técnicas multivariantes”, Dpto de Economía y Empresa, Universidad Pablo de Olavide, Sevilla, España.

[13] C. Shivert y K.E. Shirley. “LDAvis: a method for visualizing and interpreting topics”, Procedings of the Workshop on Interactive Lenaguaje Learning, Visualization, and Interfaces, páginas 63-70, 2014.

[14] Chuang,J, Manning,C.D y Heer,J. Termite: Visualization Techniques for Assessing Textual Topic Models. Stanford University, 2012.

[15] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang y D. Blei, “Reading tea leaves: how humans interpret topic models.”, presentado en Proceedings of the 22nd International Conference on Neural Information Processing Systems (NIPS'09), 2009.

[16] V. D. Blondel, J.L. Guillaume, R. Lambiotte y E. Lefebvre, “Fast unfolding of communities in large networks.”, *Journal of Statistical Mechanics: Theory and Experiment*, 2008

[17] R. Arun, V. Suresh, C.E. Veni Madhavan y M.N. Narasimha Murthy “On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. En: Advances in knowledge discovery and data mining. Springer, 391–402

[18] R. Moya. “Mapas gráficos para la visualización de relaciones en sistemas de recomendación”, Trabajo fin de máster, Dpto de Sistemas Infrmáticso, Univesidad Politécnica de Madrid, Madrid, España 2015, 95-103.

[19] Python Notebooks de la asignatura “Tratamiento de Datos” del Máster Universitario en Ingeniría de Telecomunicación, Univesridad Carlos III de Madrid. Proporcionado por Jesús Cid Sueiro.

[20] Página de la librería gensim. Fuente en línea disponible en: https://radimrehurek.com/gensim/

[21] Diapositivas Tema 0 de la asignatura “Algoritmos de Búsqueda para la Gestión de Información Multimedia” del Grado en Ingeniría de Sistemas Audiovisuales, Universidad Carlos III de Madrid.

[22] D. Ramage, S. Dumais, y D. Liebling, “Characterizing Microblogs with Topic Models” ,presentado en 4th Int'l AAAI Conference on Weblogs and Social Media, 2010,.

[23] D. Ramage, D. Hall, R. Nallapati y C. D. Manning, “Labeled LDA: A supervised topic model for credit attribution in multi-label corpora”, presentado en The Conference of Empirical Methods in Natura Language, 2009.

[24] F. Perez-Tellez, D. Pinto, J. Cardiff y P. Rosso, “Clustering Weblogs on the Basis of a Topic Detection Method.” En: Advances in Pattern Recognition. Lecture Notes in Computer Science, vol 6256. Springer, Berlin, Heidelberg, 2010.

8. Anexos

A.1 Texto del nodo C3 de la primera colección.

Fuente: http://mtjrrantsravesonmusic.blogspot.com.es/search?updated-max=2015-11-16T07:08:00-08:00&max-results=7&start=1&by-date=false

AGF & VARIOUS ARTISTS-"DEEP MYSTERIOUS TONE" (9/18) (A Deep Mysterious Tone is the third edition in AGF's poetry series, following Gedichterbe (AGF 015CD, German, 2011) and Kuuntele (AGF 017CD, Finnish, 2013]. The series investigates the history of poetry in one particular language from a female perspective. Via current music and sound art practices, the poems are reimagined in a 21st-century context. Twelve Japanese poets were selected for this issue. Three poets date from the Heian period (794-1185) and the rest are from the Meiji Restoration (when Japan opened up to the world, starting in 1868) and after. In addition to the Japanese-language poets, one piece features the sound of the Ainu, the indigenous people of northern Japan. These primarily female stories were interpreted by AGF and her contemporaries Kyoka, Tujiko Noriko, Yu Kawabata, and Ryoko Akama, fellow artists and producers she got to know at various music festivals in Europe. The only living poet represented here is Misumi Mizuki, who reads her work herself)

2. DAVE ALVIN & PHIL ALVIN-"LOST TIME" (9/18)
3. ASTRONAUTS, ETC-"MIND OUT WANDERING" (9/18)
4. BATTLES-"LA DI DA DI" (9/18)
5. RANDALL BRAMBLETT-"DEVIL MUSIC" (9/18)
6. CARTER TUTTI VOID-"F (X)" (9/18) (Featuring Throbbing Gristle's Chris Carter and Cosey Fanni Tutti (aka Chris & Cosey) plus Nik Void of Factory Floor, Carter Tutti Void first united for Mute's 2011 Short Circuit Festival in London, offering an improvisational set that was captured on the 2012 release Transverse. As that wonderful LP displayed, this May-December band were simpatico when it came to hypnotic and austere industrial music that seemed to guide itself, and now with this first proper studio album, they continue to be the great ghost in the machine. Six tracks are offered, all of them designed like sound sculptures or tape loops that slowly develop, dissolve, and then return to a robotic heart-throbbing beat. Drop the needle in a random spot and it makes much less sense than taking the full ride, and while this is a more controlled and more composed effort -- compared to their debut -- it sounds like the sum of its parts. It also brings to mind Chris & Cosey's old cohorts Coil, especially the hallucinatory landscapes of their Time Machines release. Otherwise, there is no comparison, as Carter Tutti Void create drone albums of great worth and value, leaving the other electro shaman stuck in a loop.)
7. COCOROSIE-"HEARTACHE CITY" (9/18)
8. LANA DEL REY-"HONEYMOON" (9/18)
9. DESTRUCTION UNIT-"NEGATIVE FEEDBACK RESISTOR" (9/18) (For two years now, the psychedelic Destruction Unit has been keeping the world waiting for a new album. And it s not because they ve grown up or gotten soft, rather because they ve been in the streets and in your backyards, pushing the freek agenda and immanentizing the alien-eschaton. They ve been up and down and all around this globe, battling the greedy club owners, show promoters and control pigs to bring the new american heavy underground through your back door. Now here we are, with the psychedelic Unit s second album for Sacred Bones, Negative Feedback Resistor. In the spirit of solidarity with the other revolutionary communities of our sisters and brothers, the psychedelic Unit urges you to use this album s energy, energy your speakers can hardly contain, for it s intended purpose: to break the chains which you, at the dawn of your understanding, have fastened around your hands and feet. And to see to it that the thrones of every despot erected within you are destroyed. This is crazed-psychedelic-freek-noise guerrilla warfare and these are our streets. The pigs of the law can use their system to manipulate and censor our messages. The control creeps can keep their airwaves safe and comfortable. But none of them have been able to make us turn our voices or our guitar amps down. Destruction Unit sacrificed their ears to make this album as loud of a statement as possible. Will you lend them yours?)
10. EVENING HYMNS-"QUIET ENERGIES" (9/18)
11. FENNESZ & KING MIDAS SOUND-"EDITION 1" [2-CD] (9/18)

12. ROBERT FORSTER-"SONGS TO PLAY" (9/18) (In 2008, Robert Forster, one of Australia's most respected singer-songwriters, released The Evangelist, a work that was widely regarded as his best solo album; it more than lived up to the many high points of his legendary band The Go-Betweens. In 2015, seven years after that album, fans and critics alike may wonder what's been doing since. Quite a lot, as it turns out. Producer for acclaimed albums by Brisbane bands The John Steel Singers and Halfway. An extended stint as a music critic for the Australian periodical The Monthly that was so well received that a collection of his writings was published as The Ten Rules of Rock and Roll in 2009, and was reprinted in revised and updated form in 2011. Curator and compiler of G Stands for Go-Betweens: The Go-Betweens Anthology Volume 1 (2015), the first of three lavish box-set compilations charting the career of the iconic Australian band of which he was founding member, singer, and songwriter. Still, seven years is a long time, musically speaking. Time for writing songs, time for gathering musicians, time for preparing a refreshed creative direction that took shape as Songs to Play. Ten very different Robert Forster songs recorded on a mountaintop half an hour from his Brisbane home, in an analog studio, with a troop of young musicians: talented multi-instrumentalists Scott Bromley and Luke McDonald (from The John Steel Singers), Matt Piele (drummer from Forster's touring band), and violinist and singer Karin Bäumler. The resultant album is really like nothing he's ever done before, although it retains many of the qualities listeners know from his songwriting. His work remains highly melodic, with incisive, witty lyrics attuned to real people and real lives. The surprise is in the spirit of the record, its sense of adventure and fun -- especially after the meditative reflections of The Evangelist (recorded a year after the death of The Go-Betweens co-founder Grant McLennan). Seven years have brought a bolder, wilder approach to sound, and a set of truly inspiring compositions. Pop songs. Five minute epics. A bossa nova tune. Singer-songwriter classics. Experimental, detailed production assistance from Bromley and McDonald. It's no wonder that, from the album's opening lines on the super-charged "Learn to Burn," Forster is bursting to get out and tell his story. Seven years in the making. And worth every minute.)
13. GLEN HANSARD-"DIDN'T HE RAMBLE" (9/18)
14. RICHARD HAWLEY-"HOLLOW MEADOWS" (9/18)
15. HeCTA-"THE DIET" (9/18) (Diets? Who needs 'em? Well, if you're like Buddy Hackett, or us… you do. A few years ago, I came across a 45-second monologue about trying to lose weight, performed by the comedian Buddy Hackett. It was on a 78 RPM record on the Coral Records imprint. It occurred to me that it could make an interesting dance recording given the right situation and circumstance. Equally inspirational was the book Love Saves the Day: A History of American Dance Music Culture 1970–1979 by Tim Lawrence. In reading it, I saw parallels between the dance culture of that era and the indie-rock/punk/experimental music culture. It unlocked in my mind a genre of music long dormant in memory yet an influence so prevailing in a variety of current musical genres. Woe be it unto the man whose tastes are frozen forever, for given time, space, and understanding, such things become reborn and reimagined as we search for a creative kernel of truth. I went about the realization of this idea with fellow Lambchops and electronic-minded musicians Ryan Norris (Coupler) and Scott Martin (Hobbledeions). Together we became HeCTA. Using the idea of combining the notion of "song" and elements from stand-up comedy, and electronic music and a shared love of the electronic form and its many permutations, we respectfully, playfully explored and experimented. It was to be an equal sharing of ideas, influences, and missteps. We were looking for a new way of making songs incorporating these things, to make something concise on beat. The form is not such a surprise when you consider other electronic collaborations I've been involved with: Zero 7's remix of Lambchop's "Up With People" and co-writing X-Press2's "Give It." The results became The Diet, and those songs are with us now. As HeCTA, we take our approach seriously and are respectfully aware of the great electronic music created throughout its history continuing into the present—so much so that when it came time to mix these recordings, we reached out to some of the central figures of the genre. Such greats as Morgan Geist, John McEntire, and Q all had a hand in shaping the refined sound we present to you. With invaluable creativity and engineering by Jeremy Ferguson at Battle Tapes in Nashville, we together created what we consider to be a collection of songs that move and move through you, from the dashboard to the dance floor, from Decatur to Dornburg, from Dorchester to Detroit. Suck it up, hippies. This music is our attempt to extend the boundaries of our expression and have some fun. It's not Americana, house, techno, trap, juke, or blaze. Why would it be? And like any good diet, it will be reviled then ultimately loved by all who give it a chance to work its way into their lives. We love you for trying, and we're trying to love you back. —kurtx)
16. LE BUTCHERETTES-"A RAW YOUTH" (9/18)
17. ROSE MCDOWALL-"CUT WITH THE CAKE KNIFE" (9/18) (Cut With The Cake Knife was recorded by Rose McDowall in 1988 and 1989 following the break up of her group Strawberry Switchblade. Produced with the aid of several musicians in several studios, the album features songs written for the fabled second Strawberry Switchblade album. More importantly, perhaps, it showcases the honest, direct and life-affirming songs of one of the greatest unsung songwriters of the modern pop era at a tumultuous time in her career.)
18. METRIC-"PAGANS IN VEGAS" (9/18)
19. MILD HIGH CLUB-"TIMELINE" (9/18)
20. OUGHT-"SUN COMING DOWN" (9/18)
21. POLE-"WALD" (9/18) ("Wald begins immediately, ends abruptly, and is divided into three acts of three tracks each. It is the first studio album under Stefan Betke's Pole moniker in eight years... half an eternity in the digital age -- yet the pieces on Wald seem timeless, or to have fallen from time. Stefan Betke: 'After Steingarten (SCAPE 044CD, 2007) I was on tour for two years. At the same time, Barbara Preisinger and I set up our record label ~scape. Those were two hindering circumstances that were not exactly conducive to a creative restart... I couldn't have really added anything new in the wake of

Steingarten and the dub declensions I had made in previous years.'... Over several years, long walks in the woods preceded the resumption of the production of his own material: 'Walks through the Isar valley, but also through the forests in the Alps... You have to go through life with an open mind and with extended antennae. If something strikes you and inspires you to create new music, then it will be for a reason...' For Pole, it was the forest; its spatiality... manifested, for example, in raw sounds (second act) and in psychedelic structures (third act), which sound as if they might be guitars (but are actually distorted synthetic lines)... From the tangible experience of the forest, a rather abstract question emerges: 'How can I take what I have seen or felt and make it audible?' This question becomes a narrative, a storyline. The initial story is that Pole went into a dialogue with his instruments, and the second story can be heard in the three acts of Wald. The new compositions on Wald do not deny their inheritance within the continuum of dub, yet they bring an entirely new vocabulary to Pole's sonic and spatial universe... 'If Wald had nothing to do with the world of Pole, then I would have come up with a new alter ego and produced it under a new name.' The structures, forms, and processes that Betke perceived in the forest were translated into musical structures, forms, and processes that inherently sounded like Pole. Perhaps the forest simply produces reverberations (just like the echo in the mountains!) that give rise to a bounty of thoughts. The story behind it is told in music, without the use of words -- as has previously so often been the case with Pole." --Max Dax)

22. DAVE RAWLINGS MACHINE-"NASHVILLE OBSOLETE" (9/18)
23. SWEET-"ACTION: THE ULTIMATE STORY" [2-CD] (9/18)
24. TELEKINESIS-"AD INFINITUM" (9/18) (When it came time to make Ad Infinitum, the fourth Telekinesis album, drummer/songwriter/principal architect Michael Lerner found himself in a predicament. In just under five years, he had released three fantastic records—Telekinesis! (2009), 12 Desperate Straight Lines (2011), and Dormarion (2013)—each more ambitious than the last. He had toured all over the world, shared stages with great bands, and enthralled fans of his infectious, ebullient power pop. Newly married and happily ensconced in the home studio he'd assembled in his West Seattle basement, Lerner found himself asking the question that has haunted modestly successful bands down the ages: What do you do after the rock and roll dreams you had when you were 19 have come true? "I went down to the basement," Lerner recalls, "and started playing the same chords I always play… I just felt like I'd exhausted everything I knew. I was not excited at all. I just could not make another power-pop album." While many artists have made fruitful use of vintage sounds and production techniques in recent years, Ad Infinitum is a different animal. It feels less like a time capsule and more like a time machine. In the movie version of the story, Lerner would stumble on his way down the stairs, hit his head, and wake up in 1983, and the only way he could get back to the present day would be to make a record using available instruments. Then he'd wake in 2015 to discover he'd been in his basement studio all along. And the record he'd made in that strange dream state would turn out to be Ad Infinitum, the most ambitious and assured Telekinesis release to date.)
25. VARIOUS ARTISTS-"STILL ON THE LINE: A TRIBUTE TO JIMMY WEBB" (9/18) (For the last 50 years, American popular music has been filled with the timeless songs of Jimmy Webb and we are thrilled to help continue that legacy with Still on the Line: A Tribute to Jimmy Webb. These are 12 new recordings by artists who share our passion for his songs, which we consider to be some of the most emotionally moving and skillfully crafted of all time. Part of Webb's unique talent lies in the versatility of his songs. While none of these songs are meant to replace any of their previously recorded versions, they demonstrate their strength to grab our hearts no matter the genre, voice or decade where they find themselves. All profits from the sale of "Still on the Line: A Tribute To Jimmy Webb" will go to benefit Mother Hubbard's Cupboard, a food pantry in Bloomington, Indiana. Features contributions from THE CAIRO GANG, BONNIE PRINCE BILLY, ELEPHANT MICAH, SISTER SISTER, PAN-PAN WEN, VIA VEGRANDIS, WOODEN WAND, ANDREW SLATER, POP ZEUS, and more.)

SEPTEMBER 25, 2015 (WEEK #38)

1. BANG ON A CAN ALL-STARS AND CHOIR OF TRINITY WALL STREET-"JULIA WOLFE: ANTHRACITE FIELDS" (9/25) (Haunting, poignant and relentlessly physical, Julia Wolfe's Anthracite Fields is a lovingly detailed oratorio about turn-of-the-20th-century Pennsylvania coal miners, and a fitting recipient of the 2015 Pulitzer Prize for Music. Weaving together personal interviews that she conducted with miners and their families, along with oral histories, speeches, rhymes and local mining lore, Wolfe sought to honor the working lives of Pennsylvania's anthracite region. It's not necessarily mainstream history, she told NPR shortly after she received word of winning the Pulitzer. The politics are very fascinating the issues about safety, and the consideration for the people who are working and what's involved in it. But I didn't want to say, Listen to this. This is a big political issue. It really was, Here's what happened. Here's this life, and who are we in relationship to that? We re them. They're us. And basically, these people, working underground, under very dangerous conditions, fueled the nation. That's very important to understand. Featuring the always adventurous Bang on a Can All-Stars and the renowned Choir of Trinity Wall Street, Anthracite Fields merges diverse musical styles with classical themes from the deep, ambient sweep of the opening movement Foundation (with the All-Stars Mark Stewart wrenching waves of keening sound from his electric guitar) to the high-energy rock mood of Breaker Boys.)
2. BIKINI KILL-"REVOLUTION GIRL STYLE NOW" (9/25) (BIKINI KILL is the infamously fierce, riot grrrl band featuring feminist punk pioneers KATHLEEN HANNA, TOBI VAIL, BILLY KARREN and KATHI WILCOX. Bikini Kill s first collection of work, Revolution Girl Style Now, is now released on vinyl, CD, and digital formats for the first time via the band's own Bikini Kill Records. The Revolution Girl Style Now reissue features three previously unreleased and mostly unheard tracks: Ocean Song, Just Once, and Playground. These songs feature a decidedly more grunge sound than the rest of the Bikini Kill catalog.)

3. DAVID BOWIE-"FIVE YEARS 1969-1973" [12-CD BOXED SET] (9/25) (The 12-CD box feature all of the material officially released by Bowie during the nascent stage of his career from 1969 to 1973. All of the formats include tracks that have never before appeared on CD/digitally as well as new remasters. The boxed set accompanying book, 128 pages in the CD box, will feature rarely seen photos as well as technical notes about each album from producers Tony Visconti and Ken Scott, an original press review for each album, and a short foreword by legendary Kinks front man Ray Davies. The CD boxed set will include faithfully reproduced mini-vinyl versions of the original albums and the CDs will be gold rather than the usual silver.)
4. CASPIAN-"DUST AND DISQUIET" (9/25)
5. CHVRCHES-"EVERY OPEN EYE" [TARGET EXCLUSIVE] (9/25)
6. THE DEAD WEATHER-"DODGE AND BURN" (9/25)
7. THE DEARS-"TIMES INFINITY, VOLUME ONE" (9/25)
8. DUNGEN-"ALLAS SAK" (9/25)
9. FUTUREBIRDS-"HOTEL PARTIES" (9/25)
10. GIRL BAND-"HOLDING HANDS WITH JAMIE" (9/25)
11. PATTY GRIFFIN-"SERVANT OF LOVE" (9/25)
12. JULIA HOLTER-"HAVE YOU IN MY WILDERNESS" (9/25)
13. THE INTELLIGENCE-"VINTAGE FUTURE" (9/25)
14. KASKADE-"AUTOMATIC" (9/25)
15. LOS LOBOS-"GATES OF GOLD" (9/25)
16. COURTNEY LOVE-"MISS NARCISSIST B/W RADIO KILLER" [ON GHOST RAMP 7" SINGLE] (9/25)
17. NEW ORDER-"MUSIC COMPLETE" (9/25)
18. PEACHES-"RUB" (9/25)
19. PRAYERS-"YOUNG GODS" (9/25)
20. RICKED WICKY-"SWIMMER TO A LIQUID ARMCHAIR" (9/25) ("Dayton, Ohio-based supergroup Ricked Wicky pulls off a rarely ventured and even more rarely gained three-peat with its third album—all recorded and released in the span of a year—"Swimmer to a Liquid Armchair". Serving up the same gleefully messy prog / punk / pop stew as on the previous two Ricked Wicky releases, theres a sense of assurance evident on this record that indicates a promising future.")
21. SILVERSUN PICKUPS-"BETTER NATURE" (9/25)
22. SNEAKERS-"SNEAKERS" (9/25) (Before The dB's and Let's Active, there was Sneakers! Chris Stamey, an icon of indie pop, and friend Mitch Easter began to explore recording techniques in Winston-Salem, NC, during their youth. In 1976, Chris and his band, Sneakers (including drummer Will Rigby, with appearances from Easter), released a single on Stamey's own Carnivorous Records. The sessions were engineered by Don Dixon, who would eventually produce bands like R.E.M. (with Easter) and The Smithereens. Stamey & Rigby would go on to form The dB's and Easter would reappear in Let's Active – but the Sneakers single remains vital in not only independent record history, but music in general. Omnivore Recordings is proud to reissue this seminal 7" single as an expanded CD – also available digitally. Sneakers' original six tracks are joined by five bonus tracks, including a cover of The Grass Roots' "Let's Live For Today." 9 of the tracks were available on Omnivore's 2015 10" vinyl reissue for Record Store Day Black Friday, but this new version adds "Be My Ambulance," which previously appeared on the 1978 collectible mini-album, In The Red and "Some Kinda Fool" from the out-of-print 1992 collection, Racket. Featuring photos and an essay from Scott Schinder, Sneakers will give people the opportunity to experience this genesis of jangle pop. Fans of the 80s indie scene will be beyond happy to add this release to their collection, and to experience the birth of the music they love. Whether to run to the record store or just to dance, everyone needs Sneakers.)
23. U S GIRLS-"HALF FREE" (9/25)
24. KURT VILE-"B'LIEVE I'M GOIN (DEEP) DOWN…" (9/25) (Kurt Vile does his own myth making; a boy/man with an old soul voice in the age of digital everything becoming something else, which is why this focused, brilliantly clear and seemingly candid record is a breath of fresh air. Recorded and mixed in a number of locations, including Los Angeles and Joshua Tree, b lieve I m goin down... is a handshake across the country, east to west coast, thru the dustbowl history (valley of ashes) of woody honest strait forward talk guthrie , and a cali canyon dead still nite floating in a nearly waterless landscape. The record is all air, weightless, bodyless, but grounded in convincing authenticity, in the best version of singer songwriter upcycling. In Kurt s words, I wanted to get back into the habit of writing a sad song on my couch, with nobody waiting on me. I really wanted it to sound like it s on my couch -- not in a lo-fi way, just more unguarded and vulnerable. "It s a weird, accepting, mature record, acknowledging the inherent immaturity of being a person whether father, husband , partner, adult, musician, not perfect, but compelling for its understanding ... that s life though so sad to say... --Kim Gordon")
25. WAND-"1000 DAYS" (9/25)
26. YOUTH LAGOON-"SAVAGE HILLS BALLROOM" (9/25)

OCTOBER 2, 2015 (WEEK #39)

1. AUTRE NE VEUT-"AGE OF TRANSPARANCE" (10/2)
2. THE BEVIS FROND-"EXAMPLE 22" (10/2)
3. BLITZEN TRAPPER-"ALL ACROSS THIS LAND" (10/2)
4. THE BOTTLE ROCKETS-"SOUTH BROADWAY ATHLETIC CLUB" (10/2)
5. DARK STAR-"FOAM ISLAND" (10/2)
6. DEAFHEAVEN-"NEW BERMUDA" (10/2)

7. BENJAMIN DAMAGE-"OBSIDIAN" (10/2)
8. EAGLES OF DEATH METAL-"ZIPPER DOWN" (10/2) (EODM (Eagles of Death Metal) aka Jesse Hughes ('Boots Electric') and Joshua Homme ('Baby Duck'), are back with their first new album in seven years, ZIPPER DOWN! Available worldwide on October 2, ZIPPER DOWN features 11 tracks, including the hot lead single, 'Complexity.' Recorded at Pink Duck Studios in Burbank, California, Hughes and Homme co-wrote all of the album's songs (with the exception of their cover of Duran Duran's 'Save A Prayer'), and performed all of the instruments and vocals themselves. The album is produced by Homme. 'The new album, ZIPPER DOWN, really represents to me an attitude and philosophy of life,' says Hughes. 'One should not zipper up, they should zipper down and let it all hang out.' Homme says, 'In an independent study, four out of three doctors say ZIPPER DOWN is an eargasm trapped inside a crazerbeam. And I believe them.' EODM was formed in 1998 in Palm Desert, California by best friends Jesse Hughes and Joshua Homme. Despite their band name, EODM (Eagles Of Death Metal) is not a death metal band. The story goes that a friend was introducing Joshua Homme to the death metal genre, and Homme wondered what a cross between the Eagles and a death metal band would sound like. With that, EODM was born. The band released Peace Love Death Metal in 2004, followed by Death By Sexy in 2006, and Heart On in 2008.)
9. EDITORS-"IN DREAM" [DELUXE EDITION 2-CD] (10/2)
10. GIRLS NAMES-"MY ARMS AROUND A VISION" (10/2)
11. JOHN GRANT-"GREY TICKLES, BLACK PRESSURE" (10/2) (On October 2nd, John Grant will release his third solo album 'Grey Tickles, Black Pressure' produced by John Congleton (St. Vincent, Swans) and mastered by Greg Calbi at Sterling Sound. Lyrically and musically, the 12 original songs represent Grant's most ambitious work, fusing lush strings and electronic influences with his singular wit and brutal candor. The album opens in an unsettling swirl of overlapping male voices repeating 1 Corinthians 13:4 in English and Icelandic before dissolving into fuzz, and closes with the same passage read clearly, this time by a young girl. In between, Grant's depth and range are vividly present. "Voodoo Doll" is an ode to a depressed lover, drenched in bright synths and pulsing bass lines, while "Guess How I Know" is a bonafide hell-raiser, its snarling guitar licks layered with distortion as Grant sings about a toxic yet irresistible relationship. Title track "Grey Tickles, Black Pressure" blends swelling strings and choral harmonies with Grant's darkly biting humor, as he tackles his HIV diagnosis with equal parts confusion and clarity ("I'm supposed to believe that there's some guy who'll take the pain away / There are children who have cancer, and so all bets are off / 'Cause I can't compete with that"). 'Grey Tickles, Black Pressure' follows 2013's 'Pale Green Ghosts,' which earned Grant a Best International Male Solo Artist nomination at the 2014 BRITS alongside Eminem, Justin Timberlake, Bruno Mars and Drake, and 2010's 'Queen of Denmark,' named Mojo's #1 Album of the Year. Rolling Stone calls Grant's music "richly textured, both musically and emotionally" and NPR Music's Bob Boilen says, "John Grant's songs don't mess around." Grant recently performed with the BBC Philharmonic Orchestra for BBC Radio 6 Music and just wrapped a North American tour with The Pixies. More dates TBA soon.)
12. IAMX-"METANOIA" (10/2)
13. THE ICARUS LINE-"ALL THINGS UNDER HEAVEN" (10/2)
14. JOE JACKSON-"FAST FORWARD" (10/2) (Fast Forward features four sets of four songs recorded in four different cities New York, New Orleans, Berlin and Amsterdam each with different supporting musicians, including Bill Frissell, Regina Carter, members of Galactic and Zuco103, and the 14-year old singer Mitchell Sink (from the Broadway musical "Matilda.") . In addition to 14 new Jackson songs, the album features two covers (Television s See No Evil & the classic German song Good Bye Jonny.))
15. LUMINOUS BODIES-"LUMINOUS BODIES" (10/2) (During Luminous Bodies' performance at Supernormal Festival 2014, lobster people shot foam with phallic water pistols at the audience -- an apt introduction to the band. Since then the band have continued to rampage across the land, including a standout performance at Raw Power Festival 2015, with their "scumbag lysergic racket." While it's easy to praise Luminous Bodies for their electrifying spirit, their songwriting-craft should not be overlooked. Beneath their occasionally tongue-in-cheek delivery is a band that has produced a superb debut rock record. It's an album that'll whip you up and leave you in high spirits. Luminous Bodies are Dan Hunt, Gordon Watson, Luca Zoo Franzoni, Tom Fug, and Tracy Bellaries. Recorded and mixed by Wayne Adams, Bear Bites Horse. Mastered by Sam Grant, Blank Studios. Artwork direction by Jussi Brightmore. Sleeve layout by Dirty DC. 180-gram vinyl; includes download code.)
16. MERCURY REV-"THE LIGHT IN YOU" (10/2)
17. MS JOHN SODA-"LOOM" (10/2)
18. PEARS-"LETTERS TO MEMAW" 7" SINGLE (10/2) (Letters to Memaw is a one-two punch of manic melodic hardcore from PEARS. Since taking the punk world by storm last year with their debut release Go To Prison, PEARS have been on an all-out rampage. Letters to Memaw is the first taste of new material that the band is working on for their follow up full-length due out next year. Both songs on the 7" continue to carve out PEARS unique niche, melding punchy cadence with grinding hardcore and catchy melody.)
19. SHOPPING-"WHY CHOOSE" (10/2)
20. SWIM DEEP-"MOTHERS" (10/2)
21. WAVVES-"V" (10/2)

22. ALLISON WEISS-"NEW LOVE" (10/2)

A.2 Texto del nodo G1 de la primera colección.

Here's a song from 1917:

There are smiles that make us happy
There are smiles that make us blue
There are smiles that steal away the tear dropsAs the sunbeams steal away the dew

There are smiles that have a tender meaning
That the eyes of love alone may seeAnd the smiles that fill my heart with sunshine
Are the smiles that you give to me

Lyrics by J. Will Callahan, music by Lee S. Roberts)

Here is a 2017 demonstration of the above by some people I know:From top to bottom, these photographs were made at Duke University in Durham, North Carolina; at Indian Rocks Beach, Florida; and at Orange Beach, Alabama.Sumer (with apologies to Robert Burns) is no longer icumen in; in fact, it is a-goin' out. Nevertheless, even though Mrs. RWP and I have stayed close to home, we are also smiling.

A.3 Texto del nodo F4 de la primera colección.

'Hang
In There: A DIY Covers Compilation' review"This
is a compilation that was put together by FOR THE SAKE OF TAPES to raise money for the mental health charity 'Mind', and it's easily one of the best ideas for a comp I've ever seen; all of your fave DIY / lo-fi punk bands doing covers of pop tunes from the 80s to present day. As soon as the comp started with TWO WHITE CRANES' acoustic take on BLINK 182's 'Adam's Song', followed by a great version of CHRISTINE AND THE QUEENS' ‘Tilted' by SUGGESTED FRIENDS, and WOLF GIRL doing a dreamy cover of 'I'm Not Okay' by MY CHEMICAL ROMANCE, I knew this was going to be a comp that I'd love. There are twenty-three tracks of similar brilliance here, with covers of HOLE, LE TIGRE, and BIKINI KILL by LITTLE FISTS, TOUGH TITS, and DREAM NAILS respectively, as well aspop bangers from KATY PERRY, SEAL, and THE

BANGLES by FAGGOT,EMMA KUPA & ANKLES McGHEEand FIGHTMILK. The whole comp is bloody brilliant from start to finish to be honest, but my faves are CAT APOSTROPHE doing ROBYN's 'Dancing On My Own' (it probably helps that I love ROBYN, but this cover is darn good), JESUS AND HIS JUDGEMENTAL FATHER's cover of SHANIA TWAIN'S 'That Don't Impress Me Much', DANIEL VERSUS THE WORLD's absolutely stunning version of ERASURE's 'A Little Respect' (I literally couldn't believe how good this was), and THE POTENTIALS' cover of CHER's 'Believe', pretty much because they put some of that auto-tuning from the original on the vocals in the bridge / chorus and it made me cry with laughter. So yeah, an awesome comp for a great cause that you should check out now if not sooner."https://forthesakeoftapes.bandcamp.com/album/hang-in-there-a-diy-covers-compilationxox

A.4 Texto del nodo B0 de la primera colección.

Pop and New Zealand are usually a successful cocktail, and such is the case with Kane; Strang's new single, the absolutely brilliant" My Smile Is Extinct". He really knows how to write great guitar pop jams that mix a bit of jangle with edgy power pop, and is putting out his second album,Two Hearts And No Brain,on June 30th.

A.5 Texto del nodo B1 de la primera colección.

After two great albums asWaxahatchee, Katie Cruchfield returns with a new one this summer, a record calledOut In The Storm to be released on July 14th throughMerge. "Silver"is the first single, a great 90's inspired guitar pop track with a fuller and more electric sound than her previous efforts.

A.6 Texto del nodo B6 de la primera colección.

Belgian duoSoulwaxare back with their first lp in 12 years, a record called;From Deewee;out later this month, which they say it was recorded in just 48 hours."Missing Wires"is the first advance, and it is a great, trippy and infectious electro jam.

B.1 Configuración en Gephi : Grafo con 5 tópicos para colección 1

- Layout de visualización: *Force Atlas*
- Fuerza de Repulsión = 700
- Fuerza de Atracción = 10
- Máximo Desplazamiento = 40
- Fuerza de Auto-Estabilización = 80
- Gravedad = 30

B.2 Configuración en Gephi : Grafo con 14 tópicos para colección 1

- Layout de visualización: *Force Atlas*
- Fuerza de Repulsión = 600
- Fuerza de Atracción = 50
- Máximo Desplazamiento = 40
- Fuerza de Auto-Estabilización = 80
- Gravedad = 30

B.3 Configuración en Gephi : Grafo con 4 tópicos para colección 2

- Layout de visualización: *Force Atlas*
- Fuerza de Repulsión = 1000
- Fuerza de Atracción = 45
- Máximo Desplazamiento = 40
- Fuerza de Auto-Estabilización = 80
- Gravedad = 30

B.4 Configuración en Gephi : Grafo con 3 tópicos para colección 2

- Layout de visualización: *Force Atlas*
- Fuerza de Repulsión = 1500
- Fuerza de Atracción = 60
- Máximo Desplazamiento = 40
- Fuerza de Auto-Estabilización = 80
- Gravedad = 30

Anexo C: English Summary

Introduction and motives

The Digital Revolution was the onset of the Information Age. Information storage and transmission has been boosted way beyond their previous limits. This has led to a reality in which information *management* is crucial. However, since the quantity of information available is beyond human capacity, we must look towards technology to find the means for it. This is, indeed a problem, because much information is created by humans and humans and computers don't 'speak the same language'.

A large amount of digital information is expressed in human language, formally known as natural language. As anyone can verify, when one opens any Internet explorer, it doesn't matter whether it is a website, a blog or social media, one reads *words*. Therefore, there we need computers to learn our language or, at the very least, be able to act as if they have.

The area of knowledge related to this 'human language lessons for computers' is called Natural Language Processing (NLP) and has, as its main goal, the understanding and processing of information expressed in natural language. For that, it benefits from disciplines like applied linguistics, statistical inference or automatic learning techniques.

Among NLP techniques, we find Topic Modelling (TM), a way of describing patterns that are, then, used to make predictions on the content of a given document. It first aims to describe Topic Models, that is, distributions of co-occurring words. It also looks for the algorithms that permit the use of Topic Models for making predictions.

An example of TM and one of the most common today is the Latent Dirichelt Allocation (LDA). Its goal is to gather information in form of documents and explain it in form of topics. Each of these topics is a latent variable, described as a distribution of the probability that has a number of collected words to have a certain inferred meaning.

If we think of an Internet Blog as a source of non-categorized information expressed in natural language, then we can think of them as a place where we can benefit from LDA, which is, in fact, the motive of this study. Specifically, the aim is to establish whether the content of a Blog can be faithfully represented as a combination of the mentioned latent topics and, further, shed some light into how Blogs interact with each other. We also want to point out the handicap we face, since blogs are usually made of very heterogeneous content and are written by people with different backgrounds.

Aims

In order to achieve our main and final goal (the one mentioned just above), we have set a path of subsequent aims that will lead to it. These are:

- Implement LDA topic modeling and analyze its functioning on a collection of blogs.
- Visualizing Topic Distribution
- Visualizing graphs and community detection

Background

We will now proceed to introduce some concepts that the reader will need to fully understand this work.

Blogger

A Blog is any website that has a diary structure and whose content can be whatever creation the author decides. It is also usual that readers are able to comment below each post.

There are several platforms that help people to run a blog. One is Blogger, owned by Google. All Blogs used for this study belong to Blogger.

Topic modelling with LDA

LDA modelling results in two distributions of probability: P(w | z) and P(z | θ). The former shows the probability that a certain word has to appear in any document of a given theme; the latter shows the probability that a given document has to cover a specific theme. Documents are here represented as vectors that compound a multivariate distribution of Drichlet.

Hence, LDA modelling depends on the parameters that we proceed to describe:

- **K**. Number of latent topics.
- **Alpha**, a parameter that belongs to Dirichelt distribution. It describes the *a priori* knowledge we have of how themes are represented in documents. Its value varies from 0 to 1. It is also shown as a vector of *K* dimensions.
- **Beta**, a parameter that belongs to Dirichelt distribution. It describes the *a priori* knowledge we have of how words are represented in each theme. Its value varies from 0 to 1. It is also shown as a vector of *K* dimensions.

When it comes to understanding the outcomes, the information provided must be comprehended by both computers and humans. The formers benefit of knowing the number of topics that better define the main themes of a given collection. This is why, for this approach, we used quantitative measures based on the Kullback-Leibler divergence over both the distribution of words and the distribution of documents of a given topic. For humans, however, qualitative analysis is also needed. We, therefore, used PyLDAvis to show the semantic composition of each topic generated by modelling.

Results

Visualising information obtained by Topic Modelling

Results must be understandable. This is achieved by:

- Summarizing the corpus by revealing the topics that conform it.
- Revealing relation between the analysed documents.
- Revealing relations between the latter and the topics found.

For this, the procedure is the following:

- First, topic composition is visualised using PyLDAvis library.
- Then, graphs are created by using Jensen-Shannon divergence measures.
- Finally, an analysis on centrality and community detection is run.

The first step has been already explained in the previous section. The other two are now described.

Graphs and networks analysis

The creation of this graphs aims to create a source over which we can:

- Analyse centrality measures, in order to gain more information about the whole results.
- Detect communities, in order to visualize the behaviour of each single community that conforms the whole network.

Analysis on centrality and community detection

Betweenness Centrality: it is a way to measure the centrality of a network. When we are talking about Graph Theory and Network Analysis, *centrality* is a measure that evaluates the importance of a node inside a network. It measures how many times a node appears in the shortest path that unites two other nodes. We benefit from it since it gives us a glimpse on which ones are more generic and, thus, are more related to other Topics and Blogs.

Louvain method for community detection: *communities* are sets of nodes that belong to the same networks and are more connected between them than with the rest of the nodes. *Modularity* is a function that measures the quality of the mentioned division of the network into communities. It compares how dense these connections are with how dense they would be in any well-defined network. An algorithm is used for this task. This method is based on a heuristic technique, so that modularity detection is maximized.

Conclusion

On the visualization in proposed graph, the following conclusions are established:

- Betweenness Centrality applied as a range for node size is useful for detecting nodes thar connect different communities.

- The relationships displayed in the graphs, combined with the analysis of the content of the documents, allows information extraction to characterize the different blogs and also the topics discovered.

- Community detection is useful to generate groups of similar documents and to study the relationships established between them.

On topic composition visualised using PyLDAvis library the following conclusions are established:

- Modelling heterogeneous corpus does not offer semantically interpretable information on topics composition. Therefore, the LDA model is not enough to characterize a collection of blogs.

- For characterizing collections of blogs through graph visualization it is necessary to resort to reading the content of certain relevant publications.